



## **CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS**

*Recredenciado pela Portaria Ministerial nº 1.162, de 13/10/16, D.O.U nº 198, de 14/10/2016*  
ASSOCIAÇÃO EDUCACIONAL LUTERANA DO BRASIL

Rafael Oliveira Vilarino

### FERRAMENTA PARA EXTRAÇÃO DE INFORMAÇÕES ESTRATÉGICAS DA BASE DE DADOS DO ENADE

Palmas – TO

2017

Rafael Oliveira Vilarino

FERRAMENTA PARA EXTRAÇÃO DE INFORMAÇÕES ESTRATÉGICAS DA BASE  
DE DADOS DO ENADE

Trabalho de Conclusão de Curso (TCC) elaborado e apresentado como requisito parcial para obtenção do título de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M. Sc. Fernando Luiz de Oliveira.

Palmas – TO

2017

Rafael Oliveira Vilarino

FERRAMENTA PARA EXTRAÇÃO DE INFORMAÇÕES ESTRATÉGICAS DA BASE  
DE DADOS DO ENADE

Trabalho de Conclusão de Curso (TCC) elaborado e apresentado como requisito parcial para obtenção do título de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M. Sc. Fernando Luiz de Oliveira.

Aprovado em: \_\_\_\_/\_\_\_\_/\_\_\_\_

BANCA EXAMINADORA

---

Prof. M.e Fernando Luiz de Oliveira

Orientador

Centro Universitário Luterano de Palmas – CEULP

---

Prof. M.e Fabiano Fagundes

Centro Universitário Luterano de Palmas – CEULP

---

Prof. M.e Madianita Bogo

Centro Universitário Luterano de Palmas – CEULP

Palmas – TO

2017

## **AGRADECIMENTOS**

Agradeço primeiramente a Deus por ter me dado força para superar todos os momentos de dificuldades, tenho certeza que sem ele não teria conseguido concluir esta jornada. A minha mãe por sempre ter acreditado em mim, mesmo nos momentos em que nem eu mesmo acreditava, sempre me ajudando a seguir em frente e a nunca desistir. A minha família que sempre esteve ao meu lado contribuindo para que eu pudesse concluir essa etapa da minha vida da melhor forma possível.

Agradeço aos professores, que sempre foram presentes, ajudando em todos os momentos tanto para o crescimento pessoal como profissional. A todos aqueles que contribuíram de alguma maneira para que pudesse concluir esta caminhada. Espero que Deus abençoe a vida de todos.

## RESUMO

VILARINO, Rafael Oliveira. **Extração de Informações Estratégicas da Base de Dados do ENADE 2014**. 2016. 33 f. Trabalho de Conclusão de Curso (Graduação) – Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas, Palmas/TO, 2016.

Atualmente, a quantidade de dados que uma instituição possui é tão elevada que ter uma ferramenta que sirva de apoio ao processo de tomada de decisão é de fundamental importância. Porém, muitos sistemas de informação são tidos como ineficientes por não atingirem os resultados esperados, mesmo ao aplicar modelos de decisão considerados apropriados. Sistemas eficientes para a tomada de decisões devem apresentar resultados de forma ágil, promovendo um maior envolvimento do usuário final, diminuindo conseqüentemente a possibilidade de fracasso. O foco deste trabalho visa a construção de um *Data Warehouse* (DW), no qual serão armazenados dados extraídos da base do ENADE 2014. Este DW possibilitará produzir dados referenciais que permitam a definição de ações voltadas à melhoria da qualidade dos cursos de graduação das Universidades Brasileiras por meio da aplicação de uma ferramenta OLAP (*On-line Analytical Processing*) sobre os dados armazenados de tal forma a possibilitar uma melhor análise dos dados do ENADE 2014. Para tanto, este trabalho aborda, em um primeiro momento, conceitos relacionados a *Data Warehouse*, apresentando seus modelos de dados, arquitetura, implementação, características e aplicação do processo ETL (*Extract, Transform and Load*), bem como conceitos relacionados ao processamento OLAP e à ferramenta Pentaho. Já no segundo momento, este trabalho apresenta o modelo de DW construído para trabalhar com os dados do ENADE 2014, bem como a aplicação da ferramenta Pentaho na análise dos dados, respondendo a questões que podem ser consideradas quando se deseja melhorar a qualidade de ensino de uma instituição.

**Palavras-chave:** *Data Warehouse*, ENADE, OLAP.

## LISTA DE ABREVIATURAS E SIGLAS

|        |  |
|--------|--|
| ASCII  | <i>American Standard Code for Information Interchange</i>              |
| BI     | <i>Business Intelligence</i>   |
| CEULP  | Centro Universitário Luterano de Palmas                                |
| CONAES | Comissão de Avaliação da Educação Superior                             |
| DW     | <i>Data Warehouse</i>  |
| ENADE  | Exame Nacional de Desempenho de Estudantes                             |
| ENEM   | Exame Nacional de Ensino Médio   |
| ETL    | <i>Extraction Transformation and Load</i>                              |
| HOLAP  | <i>Hybrid On-Line Analytical Processing</i>                            |
| MEC    | Ministério da Educação   |
| MOLAP  | <i>Multidimensional On-Line Analytical Processing</i>                  |
| INEP   | Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira |
| SAS    | <i>Statistical Analysis System</i>                                     |
| SGBD   | Sistema de Gerenciamento de Banco de Dados                             |
| SINAES | Sistema Nacional de Avaliação da Educação Superior                     |
| SQL    | <i>Structured Query Language</i>                                       |
| OLAP   | <i>On-Line Analytical Processing</i>                                   |
| PHP    | <i>Hypertext Preprocessor</i>  |
| ROLAP  | <i>Relational On-Line Analytical Processing</i>                        |

## LISTA DE FIGURAS

|  |    |
|--|----|
| Figura 1 – Modelo dimensional estrela. ....  | 17 |
| Figura 2 – Modelo dimensional floco de neve .....  | 18 |
| Figura 3 – Implementação top down. ....  | 20 |
| Figura 4 – Implementação bottom up.....  | 22 |
| Figura 5 – Ambiente Data Warehouse. ....   | 25 |
| Figura 6 - Metodologia.....  | 31 |
| Figura 7 - Modelagem do Data Warehouse.....  | 39 |
| Figura 8 - Arquitetura do sistema.....   | 41 |
| Figura 9 - Modelo Relacional da Base de Dados da Aplicação .....   | 42 |
| Figura 10 - Diagrama de sequência do sistema para o processo ETL.....  | 44 |
| Figura 11 - Tela inicial do sistema .....  | 46 |
| Figura 12 - Formulário de cadastro de exame .....  | 46 |
| Figura 13 - Indicando a(s) arquivo(s) de dados.....  | 47 |
| Figura 14 - Relação de arquivos do exame .....   | 47 |
| Figura 15 - Formulário de cadastro de arquivos.....  | 48 |
| Figura 16 - Lista de arquivos cadastrados .....  | 48 |
| Figura 17 - Tela para realizar a extração e carga.....   | 49 |
| Figura 18 - Configurando o Schema de um Cubo .....   | 50 |
| Figura 19 - Configuração da tabela de dimensão do cubo .....   | 51 |
| Figura 20 - Configuração da tabela de hierarquia do cubo.....  | 51 |
| Figura 21 - Configuração do level de uma hierarquia do cubo .....  | 52 |
| Figura 22 - Configuração da propriedade quantitativa do cubo .....   | 52 |
| Figura 23 - Relação entre a idade do aluno com sua nota no ENADE com os cursos com melhores desempenho.....        | 53 |
| Figura 24 - Relação entre a idade do aluno com sua nota no ENADE com os cursos com piores desempenho. ....         | 54 |
| Figura 25 - Relação entre a idade do aluno com sua nota no ENADE com os cursos com melhores/piores desempenho..... | 55 |

|  |    |
|--|----|
| Figura 26 - Relação entre a nota no ENADE dos alunos do Estado do Tocantins com o turno do curso que cursam. ....                      | 56 |
| Figura 27 – Relação entre a nota no ENADE dos alunos do Estado do Tocantins dos cursos com período integral com sua idade e sexo.....  | 57 |
| Figura 28 - Relação entre o tipo de Instituição de Ensino (pública/privado) que tiveram o melhor desempenho no ENADE.....              | 59 |
| Figura 29 - Relação entre o período de conclusão do aluno no ensino médio e seu ingresso na faculdade com seu desempenho no ENADE..... | 61 |
| Figura 30 - Download do Pentaho Server 7.0 .....   | 71 |
| Figura 31 - Tela de login do Pentaho Server 7.0.....   | 72 |
| Figura 32 - Tela inicial do Pentaho Server 7.0.....  | 73 |
| Figura 33 - Download do Mondrian Schema Workbench 3.13.0.....  | 74 |
| Figura 34 - Tela do Mondrian Schema Workbench 3.13.0 .....   | 75 |
| Figura 35 - Tela de download do drive mysql (parte 1) .....  | 75 |
| Figura 36 - Tela de download do drive mysql (parte 2) .....  | 76 |
| Figura 37 - Configurando a conexão do Mondrian Schema Workbench (parte 1) .....  | 77 |
| Figura 38 - Configurando a conexão do Mondrian Schema Workbench (parte 2) .....  | 78 |
| Figura 39 - Criando um gráfico no Saiku Analytics.....   | 79 |
| Figura 40 - Selecionando o cubo a ser trabalhado.....  | 80 |
| Figura 41 - Exibindo o gráfico .....   | 81 |
| Figura 42 - Alterando o gráfico .....  | 82 |



## **LISTA DE TABELAS**

|  |    |
|--|----|
| Tabela 1 - Dicionário de Variáveis - ENADE 2014..... | 37 |
|--|----|

## SUMÁRIO

|       |   |    |
|-------|---|----|
| 1.    | INTRODUÇÃO.....   | 12 |
| 2.    | REFERENCIAL TEÓRICO.....  | 15 |
| 2.1   | Data Warehouse.....   | 15 |
| 2.1.1 | Modelo de Dados.....  | 16 |
| 2.1.2 | Arquitetura.....  | 18 |
| 2.1.3 | Implementação.....  | 20 |
| 2.1.4 | Características.....  | 23 |
| 2.1.5 | Processo ETL.....   | 24 |
| 2.2   | Processamento OLAP.....   | 28 |
| 3.    | MATERIAIS E MÉTODOS.....  | 31 |
| 3.1   | Metodologia.....  | 31 |
| 3.2   | Linguagem de Programação.....                                       | 32 |
| 3.2.1 | PHP.....  | 33 |
| 3.3   | Software.....   | 33 |
| 3.3.1 | MySQL.....  | 33 |
| 3.3.2 | Pentaho.....  | 34 |
| 3.3.3 | Saiku Analytics.....  | 35 |
| 4.    | RESULTADOS E DISCUSSÃO.....   | 36 |
| 4.1   | Organização, Estrutura e Tipificação dos Microdados ENADE 2014..... | 36 |
| 4.2   | Questões a Serem Analisadas.....                                    | 37 |
| 4.3   | Modelo de Data Warehouse Criado.....                                | 38 |
| 4.4   | Automatização do Processo ETL.....                                  | 41 |
| 4.4.1 | Base de Dados.....  | 42 |

|  |    |
|--|----|
| 4.4.2 Diagrama de Sequência .....              | 43 |
| 4.4.3 Telas do Sistema.....                    | 45 |
| 4.5 Aplicação do Data Warehouse Proposto ..... | 50 |
| 4.5.1 Questão 1 .....                          | 53 |
| 4.5.2 Questão 2 .....                          | 55 |
| 4.5.3 Questão 3 .....                          | 58 |
| 4.5.4 Questão 4 .....                          | 60 |
| 4.5.5 Considerações .....                      | 62 |
| 5. CONSIDERAÇÕES FINAIS .....                  | 64 |
| 6. REFERÊNCIAS .....                           | 67 |

## 1. INTRODUÇÃO

O Exame Nacional de Desempenho dos Estudantes (ENADE), realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), autarquia vinculada ao Ministério da Educação (MEC), é um dos procedimentos avaliativos do Sistema Nacional de Avaliação da Educação Superior (SINAES), segundo diretrizes estabelecidas pela Comissão Nacional de Avaliação da Educação Superior (CONAES), órgão colegiado de coordenação e supervisão do SINAES.

O ENADE, componente obrigatório aos cursos de graduação, conforme determina a Lei nº 10.861/2004, consiste na avaliação do “rendimento dos concluintes dos cursos de graduação, em relação aos conteúdos programáticos, habilidades e competências adquiridas em sua formação” (INEP, 2011, Online). Sendo assim, o programa tem como objetivo avaliar (INEP, 2011, Online):

- o desempenho dos estudantes com relação aos conteúdos programáticos previstos nas diretrizes curriculares dos cursos de graduação;
- o desenvolvimento de competências e habilidades necessárias ao aprofundamento da formação geral e profissional;
- o nível de atualização dos estudantes com relação à realidade brasileira e mundial.

Anualmente, o Ministério da Educação é responsável por definir as áreas propostas pela Conaes. Em 2014, foram avaliados os cursos superiores de Ciências Exatas, Licenciaturas e áreas afins, como também os eixos tecnológicos de: Controle e Processos Industriais, Informação e Comunicação, Infraestrutura e Produção Industrial.

Normalmente, as provas são compostas por 40 questões, sendo 10 questões referentes à formatação geral e 30 relacionadas à formação específica da área, contendo questões discursivas e de múltipla escolha. As avaliações são aplicadas periodicamente aos estudantes de todos os cursos de graduação durante o primeiro (ingressantes) e último (concluintes) ano do curso.

Assim que a prova é realizada, o conceito ENADE é calculado para cada curso, tendo como unidade de observação a Instituição de Ensino Superior, o município da sede do curso e

a área de avaliação. Segundo INEP (2011, Online), a nota do curso é a média ponderada da nota padronizada dos concluintes na Formação Geral, que corresponde a 25% da nota final, e no Componente Específico que corresponde a 75%. O conceito é apresentado em cinco categorias, de 1 a 5, sendo que 1 e 2 são consideradas insatisfatórias, 3 é considerada satisfatória, enquanto 4 e 5 indicam níveis mais elevados de qualidade, na área.

O site do INEP disponibiliza os microdados gerados por avaliações, pesquisas e exames para consulta pública, incluindo os resultados do ENADE de 2004 a 2014. Tais microdados estão disponíveis para *download* em arquivos no formato ASCII (*American Standard Code for Information Interchange*). Estes proporcionam a produção de dados por instituição de educação superior, categoria administrativa, organização acadêmica, município, estado, região geográfica e Brasil.

Dentro deste contexto, este trabalho tem como finalidade analisar os resultados disponíveis do ENADE 2014, adotado como estudo de caso para este trabalho, de forma a produzir dados referenciais que permitam a definição de ações voltadas à melhoria da qualidade dos cursos de graduação por parte de professores, técnicos, dirigentes e autoridades educacionais das Universidades Brasileiras. Sendo assim, foi criado um modelo de *Data Warehouse* (DW), para onde foram extraídos e manipulados os dados em questão e, a partir de então, aplicada uma ferramenta OLAP no modelo criado, buscando se obter conhecimento. A partir das informações obtidas, apresentou-se uma análise para cada uma das questões levantadas.

A estrutura deste trabalho é organizada da seguinte forma:

- na seção 2 (dois) são apresentados os principais conceitos relacionados à *Data Warehouse*, abordando os modelos de dados, as arquiteturas, tipos de implementação e características. Em seguida são abordadas todas as fases do processo ETL, que envolve a extração, transformação e carga de dados. Por fim, relata-se como se dá o processamento OLAP;
- a seção 3 (três) descreve os materiais e métodos utilizados no desenvolvimento deste trabalho, que compreendem *softwares* e fontes de referência bibliográfica, bem como a metodologia adotada para o desenvolvimento deste;

- já a seção (quatro) apresenta os resultados obtidos durante a aplicação dos conceitos estudados, bem como as etapas envolvidas em seu desenvolvimento, com o objetivo de atingir a proposta deste trabalho;
- as conclusões obtidas durante todo o processo de desenvolvimento deste trabalho, que envolve tanto o referencial teórico quanto a aplicação dos conceitos adquiridos por meio da produção dos artefatos que compõem o produto deste trabalho, serão apresentadas na seção 5 (cinco);
- por fim, a seção 6 (seis) apresenta as referências bibliográficas utilizadas neste trabalho.

## 2. REFERENCIAL TEÓRICO

Embora uma parte considerável das instituições possuam sistemas de informação capazes de coletar e armazenar uma grande quantidade de dados em seus bancos de dados, grande parte dessas instituições tem buscado por alternativas para melhorar o planejamento, controle e gestão das informações que possuem, que é possibilitado por meio da implantação de processos de apoio à tomada de decisão e sistemas inteligentes, baseados em descoberta de conhecimento. Desta forma, a próxima seção apresenta os conceitos relacionados a *Data Warehouse* (DW), ambiente que possibilita a extração, integração, limpeza e consistência aos dados provenientes de sistemas de informação utilizados pelas instituições.

### 2.1 DATA WAREHOUSE

Nesse novo milênio a informação tem se apresentado indispensável para as organizações. Por isto, um sistema de informação integrado e eficiente é essencial para que uma organização alcance seus objetivos. Neste caso, a utilização de um *Data Warehouse* (DW) é uma opção de integração de dados.

Um ambiente de DW é formado por um conjunto de dados extraídos de uma organização que podem ser utilizadas no apoio à decisão. Segundo Inmon (1997, p. 33), DW “é um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais”. Em outras palavras, consiste em um ambiente formado por dados sólidos e concisos, resultantes de uma ou várias fontes de dados, o qual representa um novo modo de visualizar a organização no domínio estratégico e estrutural para apoio à decisão e compartilhamento de informações.

Um banco de dados relacional é utilizado para o armazenamento de dados originados de um sistema de informação. Para extrair informações de apoio à tomada de decisão, faz-se necessário implementar relatórios no sistema de informação correspondente, ficando a análise de resultados a cargo de um responsável. Já um DW é um banco de dados que guarda dados referentes aos processos do contexto que será empregado, podendo tais dados serem originados de uma ou múltiplas fontes, e não de um sistema em específico, como normalmente ocorre nos bancos de dados relacionais (INMON, 1997). O autor menciona ainda que devido ao enfoque

histórico que oferece, os dados armazenados são transformados em informações úteis para o suporte efetivo à decisão.

### **2.1.1 Modelo de Dados**

O modelo de dados tem como finalidade registrar os fatos relevantes que ocorrem no mundo real, representando de forma automatizada as necessidades de informações identificadas. Vale ressaltar que os documentos gerados devem ser utilizados como auxiliar no entendimento, não podendo ser utilizado como base para o desenvolvimento de sistemas (MACHADO, 1996).

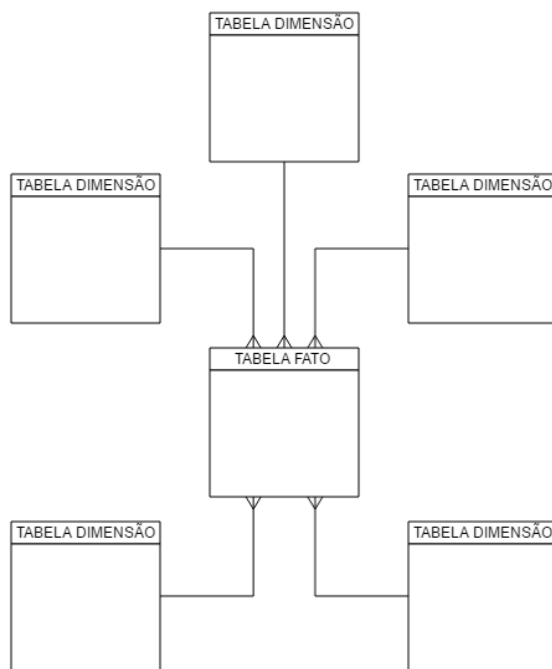
No contexto de um *Data Warehouse*, o modelo de dados determina a direção ou planejamento para a implementação de um. A elaboração deste modelo facilita a organização da estrutura e do conteúdo dos dados dentro deste DW.

Segundo Hokama et. al. (2004), a modelagem de dados pode ser descrita sobre o aspecto tradicional e multidimensional. O autor apresenta que na modelagem tradicional, objetos (clientes, fornecedores, produtos) ou transações (vendas, pedidos, notas fiscais) podem ser consideradas entidades, que se relacionam de forma direta por meio dos atributos chave. As operações estão direcionadas a dados transacionais, orientada a dados atuais que variam constantemente. Já na modelagem multidimensional, as entidades são dimensões que representam resultados para um determinado período de tempo. As entidades se relacionam indiretamente, através de outra entidade, tornando estes relacionamentos implícitos. As operações são direcionadas a dados analíticos e orientada a dados históricos estáveis.

A modelagem tradicional utiliza o modelo Entidade-Relacionamento (ER), sendo representado por entidades em conjunto com seus respectivos atributos e relacionamentos. Já na modelagem multidimensional, que utiliza bancos de dados analíticos para manipular multidimensões, são utilizados dois tipos principais de esquemas: estrela (*star schema*) e esquema floco de neve (*snowflake schema*) (HOKMA et. al., 2004). A Figura 1 representa o modelo dimensional estrela.

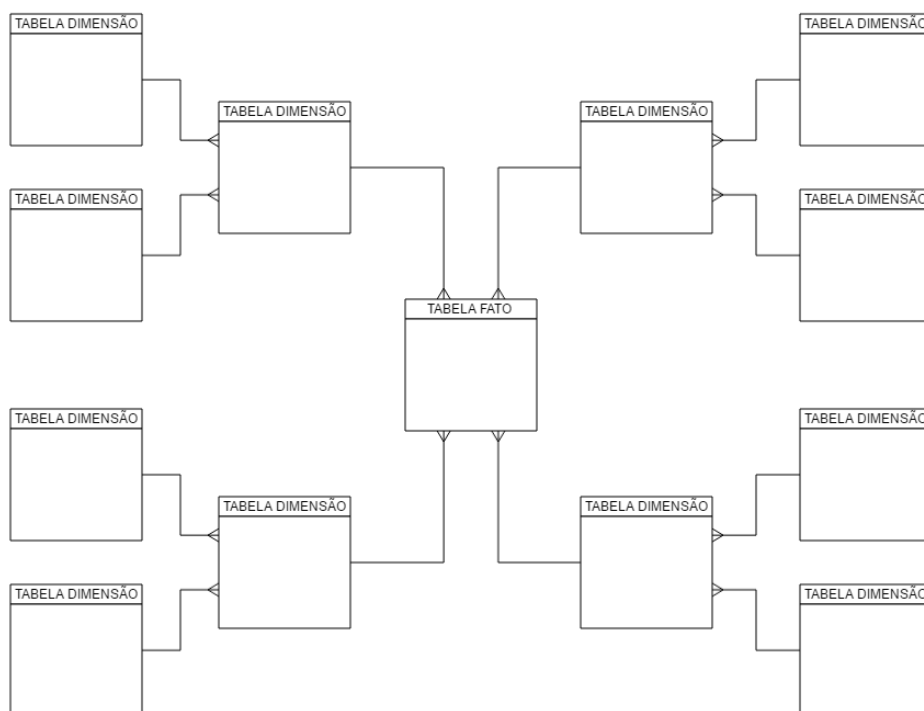


Figura 1 – Modelo dimensional estrela.



Por ser considerado a estrutura básica do modelo dimensional, o esquema estrela utiliza os mesmos componentes do diagrama entidade-relacionamento, como entidades, atributos, relacionamentos e chaves primárias (KIMBALL et. al., 1998, p. 17). Como pode ser observado na Figura 1, este modelo é formado por uma tabela central, chamada de tabela de fatos, que contém dados numéricos, sendo esta tabela relacionada com um conjunto de tabelas pequenas, cada uma chamada de tabela de dimensão. A Figura 2 representa o modelo dimensional floco de neve.

Figura 2 – Modelo dimensional floco de neve



Como pode ser observado na Figura 2, o modelo floco de neve é uma extensão do modelo estrela, onde cada uma das “pontas da estrela” torna-se o centro de outras estrelas. Assim, este modelo permite que as tabelas de dimensão se liguem entre si, bem como à tabela de fato, uma vez que estas tabelas são normalizadas. A principal vantagem na utilização deste modelo é a economia de espaço no armazenamento que este proporciona. Por outro lado, as consultas são mais elaboradas devido à complexidade sobre o número de tabelas relacionadas. A próxima seção aborda a arquitetura de um *Data Warehouse*.

### 2.1.2 Arquitetura

Segundo Singh (2010), a arquitetura de um *Data Warehouse* objetiva determinar os processos que permitirão inserir, gerenciar e exibir os dados, por meio de um conjunto de normas que proporcione uma estrutura para o projeto.

A arquitetura de dados para um projeto de DW pode ser dividida em arquitetura geral dos dados ou topologias e funcional. A arquitetura geral dos dados ou topologias possibilita

identificar e entender como os dados fluem através do DW. As arquiteturas de dados mais comuns são:

- Centralizada: caracterizada por um único DW que atende a toda a comunidade de usuários;
- <sup>1</sup>Data Marts dependentes: é composto por vários Data Marts (DM) ligados a um DW. Cada DM tem um escopo de dados limitados orientados a um tema específico do negócio, podendo os usuários conectarem tanto aos seus DM quanto ao DW;
- Data Marts independentes: é caracterizado por permitir a ligação dos usuários aos seus respectivos DM, fornecendo as informações necessárias. Uma vantagem a ser destacada referente a esta arquitetura é o fato desta tornar o desenvolvimento mais ágil, bem como baixo custo e controle local ao invés do centralizado;
- Data Warehouse Distribuído: é composto por vários DW interligado por meio de uma rede com forte suporte a processamento distribuído.

Quanto a arquitetura funcional, o DW é construído a partir de duas partes distintas: área interna e externa. A área interna é responsável por realizar as aquisições de dados a partir dos sistemas tradicionais ou de outras formas. Neste momento, o dado é identificado, copiado, formatado e preparado para ser carregado no repositório de dados do DW, que pode ser administrado através de banco de dados relacionais ou multidimensionais. Os dados que foram extraídos de fontes externas são armazenados na área de *Staging*, para então serem tratados, limpos e carregados ao DW.

Pereira (2000) indica ainda como partes da área interna:

- Carga de dados: tem como objetivo permitir o armazenamento dos dados transformados no servidor de apresentação;
- Controle dos dados organizados: permite o monitoramento sobre o fluxo de dados por meio dos metadados;
- Gerenciamento dos recursos da área interna: tem como finalidade fazer com que o DW volte a trabalhar normalmente após a ocorrência do problema.

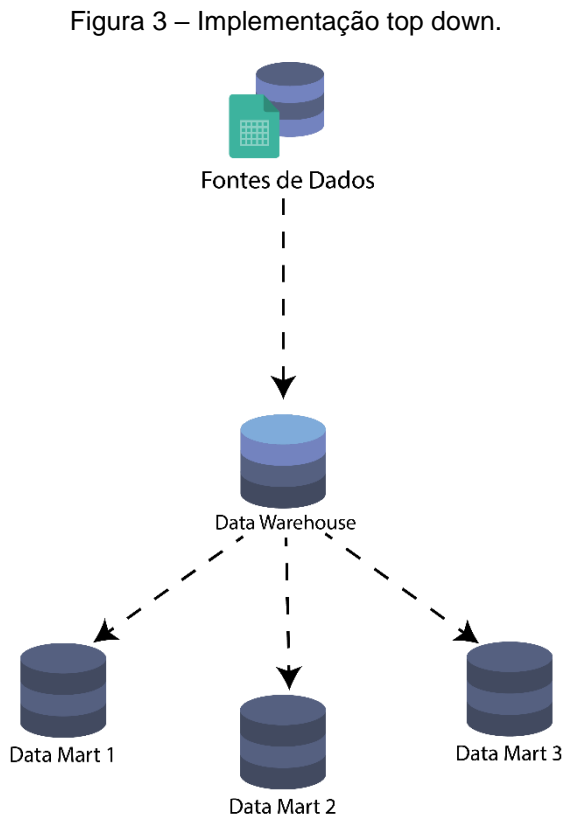
---

<sup>1</sup> Data Marts consistem em um subconjunto de informações dos *data warehouse*, que tem como finalidade aumentar a velocidade da consulta de informações, segmentando-o para um determinado assunto ou área (ANZANELLO, 2002).

A segunda área que compõe a arquitetura funcional de um DW, chamada área externa, consiste na interface do usuário com o sistema, ambiente que o usuário trabalha, sobretudo por meio de consultas (PEREIRA, 2000). Esta área é composta pelo servidor de apresentação, onde os dados nativos da parte interna ficam disponíveis para os usuários finais; e por ferramentas de acesso a dados e geradores de relatórios. A próxima seção aborda a implementação de um *Data Warehouse*.

### 2.1.3 Implementação

Há três formas de se implementar um *Data Warehouse*: *top down*, *bottom up* e a combinação das técnicas *top down* e *bottom up*. Ao escolher qual forma de implementação adotar, deve-se levar em consideração os recursos disponíveis, a escolha da arquitetura, as necessidades de acesso a dados, entre outras variáveis. A Figura 3 apresenta o desenvolvimento *top down*.



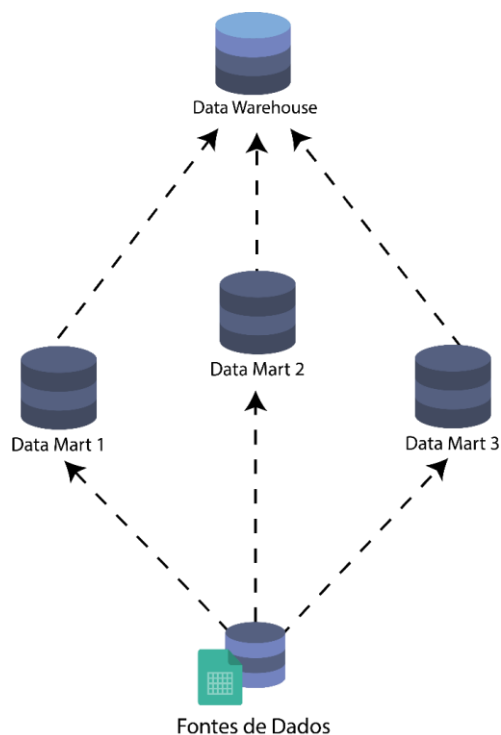
Fonte: Elaborada pelo autor.

Como pode ser observado na Figura 3, na implementação *top down*, os *Data Marts* são derivados do *Data Warehouse*. Esta forma de implementação requerer a definição do modelo de dados, uma vez que antecede a implementação, bem como necessita de uma visão geral de toda a organização (MAZZOLA, 2002).

Para que seja possível tomar decisões concernentes à utilização de fontes de dados, segurança, estrutura, qualidade, padronização e modelo de dados é imprescindível o envolvimento do pessoal de diversas áreas da empresa nesta fase. Sendo assim, pode-se dizer que esta forma de implementação exige um planejamento maior e adequação sobre o projeto nas fases iniciais, consumindo na maioria das vezes muito tempo antes de apresentar resultados, bem como um alto custo. Entretanto, há uma definição de dados e regras de negócio mais consistente e uma estruturação do DW mais global, tendo em vista que as regras de negócios são claramente definidas desde o início do projeto.

Atualmente, a abordagem *top down* tem sido pouco utilizada por necessitar de um Sistema de Informação centralizador que gerencie todos os recursos de *hardware*, aumentando seu tempo de implementação, além de dificultar e tornar maiores seus custos (MAZZOLA, 2002). A Figura 4 apresenta o desenvolvimento *bottom up*.

Figura 4 – Implementação bottom up



Como pode ser observado na Figura 4, no desenvolvimento *bottom up* o *Data Warehouse* é derivado dos *Data Marts*. Esta forma de implementação é a mais utilizada atualmente, por apresentar resultados imediatos, tendo desta forma um retorno do investimento mais rápido, se comparado ao *top down* (MAZZOLA, 2002). O que diferencia esta abordagem da apresentada anteriormente é que a *bottom up* é baseada na construção de *Data Marts*, pequenos *Data Warehouses*, que podem ser construídos antes ou de forma paralela ao DW global. Os dados para este tipo de abordagem são retirados tanto do DW global quanto de sistemas operacionais ou fontes externas de dados. Por isso, a medida que o tempo vai passando a sua implementação, que começa com um *Data Mart*, pode ser expandida. Até o momento, este tipo de implementação tem se mostrado bem mais satisfatório que os da abordagem *top down*.

Como desvantagem de utilização da *bottom up* é possível mencionar retrabalho quando são necessárias modificações em partes já implementadas ou quando novas áreas são atendidas, ocasionando a possibilidade de ocorrer redundância de dados e inconsistência entre os *Data Marts*, comprometendo desta forma sua eficiência (MAZZOLA, 2002). O autor ressalta que esta desvantagem se dá ao fato da *bottom up* não exigir que uma infraestrutura global seja

erguida anteriormente à implementação, o que não ocorre na *top down*. Portanto, é de suma importância que sejam maiores cuidados em seu planejamento, monitoramento e estabelecimento de diretrizes que influenciarão diretamente no seu melhor desempenho.

Por fim, a última forma de implementação consiste na combinação das técnicas *top down* e *bottom up*, mesclando a maior quantidade de características positivas de cada uma das técnicas. Sendo assim, requer um cuidadoso monitoramento do processo, bem como acompanhamento e gerenciamento de novos possíveis requerimentos (MAZZOLA, 2002).

#### **2.1.4 Características**

Inmon (1997) destaca as seguintes características de um *Data Warehouse*: ser orientado por assuntos, integrado, variável no tempo, não volátil, garantia quanto a confiabilidade dos dados e granularidade, as quais serão explanadas a seguir:

- orientação por assunto: significa dizer que apenas as informações relacionadas ao interesse do projeto serão empregadas, diferente de quando se trabalha com o conceito de *Big Data* onde toda e qualquer informação é aplicada no projeto.
- integração: os dados passam por um processo de integração ao serem extraídos. Por exemplo, quando dois ou mais sistemas tem seus dados extraídos para um DW, seus atributos iguais sofrem uma integração de valores.
- variação no tempo: resultados oferecidos pelo DW são específicos de um determinado período como, por exemplo, do período que foram extraídos de outros sistemas.
- não volatilidade: geralmente os dados são carregados em uma única vez e não sofrem atualizações, e quando acontece uma nova carga os dados existentes não são alterados.
- localização: na busca por elevar ao máximo o poder de processamento e agilizar a pesquisa dos dados, a base de dados de um DW é integrada, ou seja, os dados podem ser armazenados fisicamente em um único local. Além disso, podem ser armazenados de forma distribuída (*Data Marts*), sendo classificados por áreas de interesse; bem como por níveis de detalhes, em que as unidades de dados são mantidas no DW.
- credibilidade dos dados: a integração e consistência dos dados são garantidos a partir dos conjuntos de base de dados, processos de entrada, metadados e informações referente a

procedência dos dados. Vale ressaltar que todos esses aspectos são de fundamental importância para o sucesso de qualquer projeto.

- granularidade: se refere ao nível de detalhe ou resumo dos dados existentes em um DW. Quanto maior o nível de detalhes, menor o nível de granularidade, o qual interfere diretamente o volume de dados armazenados e o tipo de consulta que pode ser respondida (INMON, 1997).

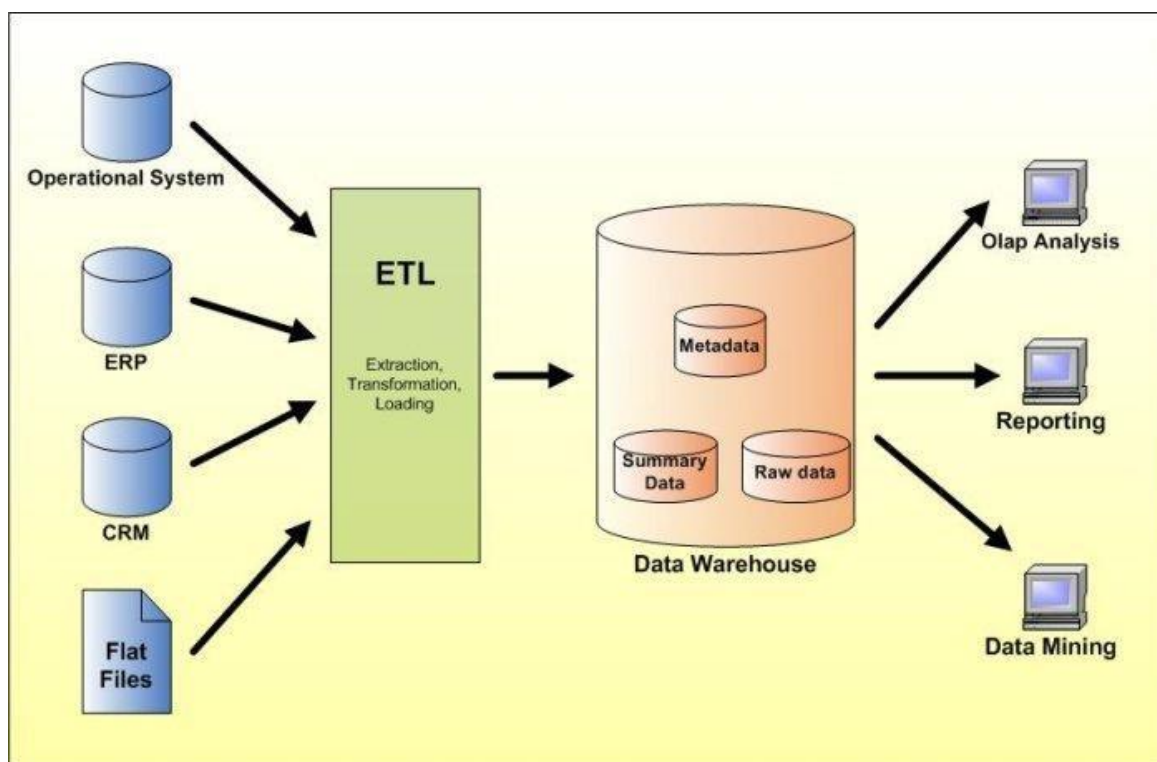
Visto as características de um *Data Warehouse*, a próxima seção apresentará conceitos relacionados ao processo ETL.

### **2.1.5 Processo ETL**

Em um ambiente de *Data Warehouse*, o processo é iniciado a partir da extração de dados de sistemas operacionais e de fontes externas para, logo em seguida, serem integrados e transformados para, finalmente, serem carregados no DW. Vale ressaltar que a fase de transformação consiste na limpeza, eliminação, combinação, validação, consolidação, agregação e sumarização dos dados. Sendo assim, um DW contém dados sumarizados, históricos e detalhados para suportar a tomada de decisões táticas e estratégicas. A Figura 5 representa o fluxo realizado em um ambiente DW até a disponibilização dos dados aos usuários.



Figura 5 – Ambiente Data Warehouse.



Fonte: (REBOUÇAS, 2010, Online).

Como pode ser observado na Figura 5, inicialmente tem-se diferentes tipos de entrada de dados (representado na figura em azul), os quais podem estar em diferentes sistemas de gerenciador de banco de dados e mainframes, bases de dados distintas, normalizados diferentemente também. Ao passar pelo processo ETL (representado na figura em verde), considerado um dos momentos mais críticos da realização do projeto, os dados são armazenados no DW de forma que seja possível recuperar os dados (*raw data*), os dados sobre os dados (*metadata*) e o agrupamento dos dados (*summary data*). A partir desse ponto, o DW já está preparado para alimentar os sistemas de apoio a decisão como ferramentas OLAP, relatórios e *Data Marts*.

### 2.1.5.1 Extração

Para se obter dados para o ambiente de *Data Warehouse*, deve-se iniciar o processo pela extração de dados. Este primeiro passo envolve o entendimento das fontes de dados e a cópia das áreas necessárias para a área de transformação de dados, que serão trabalhadas

posteriormente. Como visto antes, os dados a serem armazenados na maioria dos DW são originados de várias fontes diferentes e independentes, sejam tais fontes bases de dados de sistemas tradicionais, planilhas Excel, dentre outras.

A etapa mais complexa da extração é definir quais dados extrair e determinar quais tipos de filtros aplicar, consumindo desta forma bastante tempo no processo de construção do DW. Hokama et. al. (2004) apresentam duas possíveis formas para se realizar a extração de dados. A primeira se refere à implementação de um sistema que consulte a base de dados do sistema fonte gerando arquivos com os dados desejados. Já a segunda forma se dá por meio da utilização de ferramentas próprias para extração de dados, as quais serão responsáveis pela leitura e geração de arquivos com os dados definidos.

#### ***2.1.5.2 Transformação***

Concluída a fase de extração de dados, é necessário converter os dados obtidos para um formato adequado de forma que possam ser utilizados na carga do DW. Esta fase, chamada de transformação, envolve uma série de atividades, definidas conforme a necessidade e situação do projeto. Segundo Hokama et. al. (2004), as atividades mais comuns na transformação de dados são:

- **Limpeza:** esta atividade tem como objetivo deixar os elementos de dados uniformizados (dentro de formatos padrões), únicos (não duplicados), corretos e consistentes de forma a espelhar a realidade. Para isso, faz-se necessário corrigir o uso incorreto ou inconsistente de códigos e caracteres especiais, solucionar possíveis problemas de conflito de domínios, corrigir valores errados ou duplicados, e tratar dados perdidos;
- **Eliminação:** consiste em identificar campos e dados provenientes de sistemas legados que não são úteis para o DW, de forma que estes sejam descartados;
- **Combinação:** é executada quando os mesmos valores de chaves das fontes de dados representam registros iguais ou complementares;
- **Desnormalização e normalização:** o objetivo principal desta atividade é reunir as hierarquias de dados dentro de uma única dimensão, de forma desnormalizada. Entretanto,

nos casos de haver dados completamente desnormalizados dentro de arquivos texto, deve-se normalizá-los;

- Cálculos, derivação e alocação: durante o processo de levantamento de requisitos são identificadas as regras de negócio, as quais passarão pelo processo de transformação. Essas transformações envolvem manipulação de textos, aritmética de data e hora, entre outros conjuntos de funções.

Concluído o processo de transformação, os dados estão aptos para serem carregados no DW.

### **2.1.5.3 Carga**

A carga de dados no DW, atividade realizada após os dados terem sido extraídos e transformados, possui uma complexidade, exigindo que sejam levados em consideração os seguintes fatores (HOKAMA, 2004):

- Integridade dos dados: consiste em verificar se os dados existentes na tabela da chave estrangeira estão de acordo com a chave primária;
- Tipo de carga a ser realizada: há a possibilidade de se realizar a carga de forma incremental ou total, que consiste em excluir os dados existentes e incluí-los novamente. O que diferencia o tipo de carga incremental da carga total é que a primeira é realizada na tabela de fatos e a segunda nas tabelas de dimensão;
- Otimização do processo de carga: consiste em invocar um conjunto de técnicas para otimizar o processo de carga a partir dos bancos de dados ou registrá-lo em *scripts* por meio da utilização de ferramentas referente a área de organização de dados. Como exemplo dessas técnicas pode-se citar: evitar a geração de *log* durante o processo, criar índices e agregar dados;
- Suporte completo ao processo de carga: consiste em eliminar e recriar índices e particionamento físico de tabelas e índices, oferecendo desta forma suporte as exigências do serviço antes e depois da carga atual.

Assim, uma vez que é realizada a extração de dados, sendo estes transformados e carregados no DW, o processo ETL é concluído, estando hábil para a aplicação de uma ferramenta OLAP (*On-Line Analytical Processing*).

## 2.2 PROCESSAMENTO OLAP

O *On-line Analytical Processing* (OLAP) é composto por atividades gerais de consulta e apresentação de dados numéricos e textos derivados de um DW, bem como as formas específicas de consulta e apresentação que são exemplificadas por uma grande quantidade de ferramentas OLAP (KIMBALL et. al., 1998, p. 21). Em outras palavras, sistemas OLAP auxiliam os especialistas da área a resumirem informações por meio de conferições, espectros personalizados, importância dos dados e apreciação histórica em diversos panoramas referente a situações variadas e não uniformes.

Segundo Kimball et. al. (1998, p. 21), o processamento OLAP pode ser representado por abordagens distintas, sendo elas:

- *Relational On-line Analytical Processing* (ROLAP): o modelo relacional armazena os dados e o gerenciador de banco relacional processa as consultas, atribuindo características dimensionais ao banco de dados relacional. O Sistema de Gerenciamento de Banco de Dados (SGBD) relacional é composto pelas tabelas de sumário, as quais são integralmente deriváveis e, quando necessário, seus índices criados de forma automática. Apesar de permitir a utilização do padrão *Structured Query Language* (SQL), a consulta tem baixo desempenho;
- *Multidimensional On-line Analytical Processing* (MOLAP): é composto essencialmente de um banco de dados multidimensional, por meio de um conjunto de interfaces de usuário, aplicações e banco de dados, com tecnologia proprietária. Armazenam seus dados em um cubo com diversas dimensões que são trazidos para o servidor OLAP, sendo organizados em arranjos com alto grão de agregação. Quando comparado ao ROLAP, permite consultas mais rápidas;
- *Hybrid On-line Analytical Processing* (HOLAP): é composto por uma nova tecnologia que consiste na junção das principais características das abordagens ROLAP e MOLAP, com

a finalidade de resolver os principais problemas encontrados em cada uma das abordagens. A retenção dos dados é realizada por um banco de dados SGBD e as agregações por um MOLAP. Sendo assim, tem-se a alta performance do MOLAP e a escalabilidade do ROLAP. Como desvantagem pode-se destacar o fato da consulta ser mais lenta do que o modelo MOLAP quando esta é feita sobre dados básicos.

Dentre as ferramentas que auxiliam no processo OLAP, encontram-se as destacadas a seguir:

- Cognos: oferece uma vasta capacidade analítica, possui uma arquitetura totalmente agregada na interface do usuário final, possui capacidade de análises preditivas e estatísticas adequadas, e possui uma versão mais densa e com preço menor. Entretanto, seu desempenho não é considerado bom, sua implementação é mais complexa se comparada a dos concorrentes, além do custo da licença por usuário ser considerado alto e acima da média, bem como o suporte ao cliente ser de baixa qualidade (COGNOS IBM, 2016, Online).
- MicroStrategy: suporta um grande volume de dados com boa performance, possui um alto nível de satisfação por parte de seus clientes, e oferece boa produtividade por desenvolvedor. Por outro lado, a utilização de algumas funcionalidades é complexa (MICROSTRATEGY, 2016, Online).
- SAS: sua plataforma é focada principalmente em técnicas analíticas avançadas, e oferece uma ampla variedade de funcionalidades específicas não triviais. Contudo, apresenta alta dificuldade de implementação, possui diversas interfaces inconsistentes em uma suíte, bem como usabilidade limitada (SAS, 2016, Online).
- Oracle *BusinessIntelligence*: suporta o maior número de usuários e volume de dados, a mais ampla variedade de funcionalidades e a maior capacidade de carga de trabalho analítica se comparado aos concorrentes. Todavia, apresenta baixa aceitação no que se refere as tecnologias preditivas (PORTIFÓLIO SAP BUSINESSOBJECTS, 2016, Online).
- Pentaho: oferece gestão e distribuição de informes e painel de controle sobre plataforma de código aberto, integra com diferentes fontes de dados e aplicativos que utilizam padrões abertos de forma simples, é inteiramente personalizável, oferece várias opções de formatos

de arquivos de saída, bem como roda em vários sistemas operacionais. No entanto, possui uma documentação fraca, acarretando a necessidade de o usuário possuir um alto conhecimento técnico (GOLDWIN, 2016, Online).

Com o objetivo de facilitar o acesso aos dados, tais ferramentas permitem aumentar ou diminuir o nível de detalhamento das consultas às tabelas de dimensão e fato, por meio dos recursos a seguir:

- Drill-up e drill-down: uma vez que agrupa e desagrupa dados de forma progressiva, há a possibilidade de navegar entre níveis de agregação (POE & KLAUER & BROBST, 1998);
- Pivoting: a partir da agregação de duas dimensões há a possibilidade de se comparar o resultado (DBMiner, 2016);
- Slice and dice: *slice* é responsável por extrair informações sumarizadas em um cubo de dados e *dice* por extrair de um subcubo ou da intersecção de vários *slices*. Sendo assim, esse recurso permite a visualização dos dados de diferentes pontos de vistas, o que reduz a dimensionalidade dos dados.

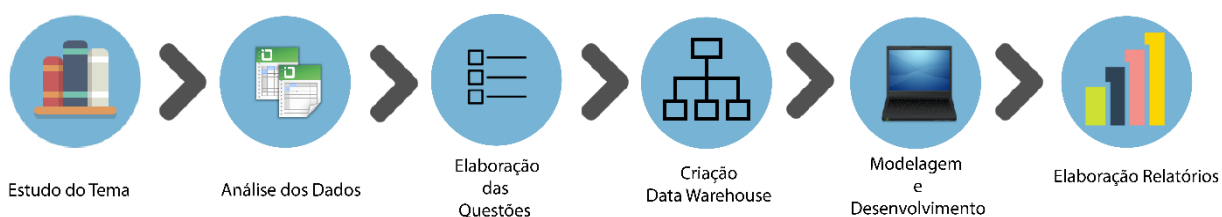
### 3. MATERIAIS E MÉTODOS

Nesta seção são apresentados os materiais utilizados e a metodologia adotada para a elaboração do presente trabalho.

#### 3.1 METODOLOGIA

Propôs-se para este trabalho uma pesquisa aplicada, de natureza quantitativa, com objetivo metodológico exploratório. O desenvolvimento deste trabalho teve início a partir de reuniões com o Prof. M. Sc. Fernando Luiz de Oliveira com a finalidade de definir uma proposta para o mesmo. Sendo assim, definiu-se que a proposta deste trabalho consiste na construção de um *Data Warehouse* (DW) para trabalhar com os dados do ENADE 2014 a fim de realizar uma análise de dados por meio da utilização de uma ferramenta OLAP sobre o DW construído. A partir desse processo é possível obter informações que possam ser utilizadas na definição de ações voltadas a melhoria da qualidade dos cursos de graduação das Instituições de Ensino. A metodologia utilizada no desenvolvimento deste trabalho é representada pela Figura 6.

Figura 6 - Metodologia



Assim que a proposta foi elaborada, iniciou-se uma pesquisa em diversas fontes de informações, com o intuito de encontrar materiais tais como, monografias, teses, artigos científicos, etc. para que se pudesse adquirir conhecimento sobre o tema abordado.

Passada a fase de estudos sobre os conceitos relacionados a este projeto, realizou-se um levantamento dos dados que serão utilizados na geração de conhecimento, selecionando-se dados relacionados ao desempenho do aluno especificamente em seu curso e em áreas gerais, de forma a relacionar e agrupar tal desempenho às características pessoais, geográficas e

acadêmicas do mesmo. Tais dados foram escolhidos com o objetivo de determinar grupos (sexo, idade, turno, estado, região, curso).

Concluída a seleção de dados, criaram-se questões que serão respondidas a partir da utilização do DW, sendo elas:

1. Qual a relação entre a idade do aluno com sua nota no ENADE com os cursos com melhores/piores desempenhos?
2. Qual a relação entre a nota no ENADE dos alunos do Estado do Tocantins com o turno do seu curso?
3. Qual o tipo de Instituição de Ensino (pública/privado) que tiveram o melhor desempenho no ENADE?
4. Qual a relação entre o período de conclusão do aluno no ensino médio e seu ingresso na faculdade com seu desempenho no ENADE?

Tais questões foram elaboradas a partir dos dados que permitem avaliar que elementos interferem/constroem o desempenho dos estudantes. Tendo os dados e questões a serem trabalhadas em mãos, definiu-se o modelo de dados do DW e que sua implementação será da forma *top down*.

Tendo definido tal modelo, implementou-se um sistema para automatizar o processo de extração de dados para o DW. Vale ressaltar que a cada nova aplicação do ENADE, poderá ser feita uma nova carga no DW implementado, mantendo desta forma um histórico dos resultados do exame a partir do ano de 2014. Assim que a implementação do sistema responsável pela extração de dados para o DW foi concluída, integrou-se o DW ao sistema.

A partir de então, aplicou-se a ferramenta OLAP Pentaho, chegando-se, desta forma, ao resultado deste trabalho. A partir dos resultados obtidos, realizaram-se análises sobre os mesmos, a fim de identificar possíveis pontos que podem interferir positiva ou negativamente no aprendizado dos alunos durante a graduação.

### 3.2 LINGUAGEM DE PROGRAMAÇÃO

Como mencionado, foi desenvolvido um sistema para realizar o processo ETL no DW. Adotou-se a linguagem de programação PHP, a qual será apresentada na próxima seção.



### **3.2.1 PHP**

Como apresentado, o *Hypertext Preprocessor*, PHP, foi adotado como a linguagem de programação a ser utilizada na implementação do sistema responsável por realizar a extração, transformação e carga de dados no DW.

O PHP, é uma linguagem de programação formada por um conjunto de scripts escritos em linguagem C. Esta linguagem foi criada para ser utilizada em páginas dinâmicas Web. Segundo Oglio (2015), estima-se que mais de 80% dos servidores Web existentes utilizam o PHP.

Sousa (2009) menciona que uma grande vantagem em utilizar a linguagem PHP deve-se a possibilidade de se deparar com inúmeros scripts prontos e testados na Internet, bem como o fato de haver vários fóruns de discussão sobre o assunto. Além disso, tem-se a liberdade de optar por utilizar a programação estrutural ou orientada a objetos, bem como o sistema operacional e servidor Web a ser utilizado no desenvolvimento de sistemas (BUYENS, 2002).

## **3.3 SOFTWARE**

Dentre os *softwares* utilizados na implementação deste trabalho encontram-se o banco de dados MySQL e a ferramenta Pentaho, os quais serão apresentadas nas próximas seções.

### **3.3.1 MySQL**

Como apresentado, o MySQL foi adotado como o SGBD responsável por gerenciar os dados do DW construído. Sendo assim, o MySQL é um Sistema de Gerenciamento de Banco de Dados (SGBD) relacional que utiliza a linguagem padrão SQL. Entre os bancos de dados de código-fonte aberto, o MySQL é o mais popular, sendo largamente usado em aplicações para a Internet.

Niederauer & Prates (2006) destacam as seguintes características para o SGBD em questão: velocidade, escalabilidade e confiabilidade. Os autores listam ainda as seguintes vantagens em se utilizar o MySQL: número ilimitado de acesso por usuários simultâneos,

capacidade de manipulação de tabelas com mais de 50.000.000 registros, alta velocidade de execução de comandos, bem como fácil e eficiente controle de privilégios de usuários. Vale ressaltar a excelente performance do MySQL quando utilizado em sistemas de informação desenvolvidos na linguagem de programação PHP.

### 3.3.2 Pentaho

A Pentaho Corporation é uma empresa de origem norte-americana, sediada na Florida, EUA. Esta empresa desenvolveu em 2004 o Pentaho Business Intelligence, uma plataforma completa de *Business Intelligence* (BI) desenvolvida em Java, distribuída e implantada como *open source*, apresentando grande flexibilidade e independência de plataformas, alta confiabilidade e segurança a um custo mínimo de implantação e manutenção (BOUMAN & DONGEN, 2009, tradução nossa).

A ferramenta Pentaho é utilizada no processo OLAP para elaboração de relatórios, *dashboards*, mineração de dados e ETL. Esta ferramenta engloba a integração de dados com análise de negócios em uma plataforma moderna que reúne especialistas da área da Tecnologia da Informação e usuários de negócios para acessar, visualizar e explorar todos os dados que impactam nos resultados do negócio (PENTAHO, 2010, tradução nossa).

Dentre as vantagens encontradas na ferramenta Pentaho, pode-se destacar (PENTAHO, 2010, tradução nossa):

- Usabilidade: oferece componentes prontos para serem utilizados;
- Habilidades: utiliza padrões a fim de facilitar possíveis integrações;
- Orientação a ferramentas: sistema completamente integrado, concentrado em processos e soluções;
- Customização/Extensão: disponibiliza componentes reutilizáveis, bem como permite que toda infraestrutura, componentes, mecanismos e interfaces sejam modificados, disponibilizando seu código-fonte;
- Foco em relatórios e análise: concentra-se em processos e *workflow*, além de ser direcionada a regras de negócios;
- Auditoria: possui robustos processos de rastreamento.

De forma geral, o Pentaho oferece diversas ferramentas de análise de dados, monitoramento de indicadores e *data mining*, oferecendo subsídios para que as instituições/empresas revolucionem a utilização da informação gerencial, alcançando proveitos expressivos de eficiência e eficácia.

Destaca-se que, embora ofereça recursos para as áreas mencionadas, o Pentaho será utilizado neste trabalho para elaborar relatórios a partir de um processo de análise do DW com os dados do ENADE. Os Apêndices A e B apresentam o processo de instalação e configuração das ferramentas utilizadas neste trabalho. Vale ressaltar que o sistema operacional utilizado para executar os passos especificados no presente trabalho foi o Linux Mint 18 Cinnamon 64-bit, porém, existem versões para outros sistemas operacionais. Outra restrição é que algumas ferramentas aqui abordadas necessitam que o JAVA esteja instalado e devidamente configurado.

### **3.3.3 Saiku Analytcs**

O Saiku Analytcs é uma ferramenta de código aberto, disponível como plugin para o Pentaho BI Server. Para facilitar a utilização do recurso de Cubos OLAP com uma experiência simples para o usuário final, o plugin utiliza o Pentaho Analisys, Mondrian. Por este motivo, é considerada facilmente incorporável, extensível e configurável.

O Saiku disponibiliza uma ferramenta analítica, baseada em tecnologia web, que oferece aos usuários a disposição de avaliar e elaborar relatórios, de forma simples e ágil, de acordo com informações contidas em repositórios multidimensionais (BATISTA, 2014).

## 4. RESULTADOS E DISCUSSÃO

Os conceitos estudados sobre Data Warehouse e Processamento OLAP proporcionaram uma fundamentação teórica sobre formas de construir um ambiente formado por informações consistentes e sucintas, resultantes de uma ou mais fontes de dados, proporcionando uma nova maneira de visualizar a organização no domínio estratégico e estrutural quanto ao suporte à decisão e compartilhamento de informações.

Como mencionado anteriormente, os dados trabalhados neste projeto referem-se aos resultados do exame ENADE referente ao ano de 2014, sendo adotado como estudo de caso. Estes dados estão disponíveis em planilhas para *download* no site do INEP. Com o intuito de criar um ambiente apropriado para subsidiar a tomada de decisão, esta seção apresentará informações retiradas da base de dados do exame ENADE referente ao ano de 2014 e considerações sobre a análise realizada sobre estes dados, a fim de avaliar o desempenho acadêmico dos estudantes no que se refere aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação, bem como a outras áreas do conhecimento relacionadas à realidade brasileira.

Para tanto, serão apresentados a organização, estrutura e tipos dos dados referente ao exame disponível para análise. A partir de então, serão apresentadas as questões que foram levantadas, o DW criado, o sistema desenvolvido para realizar todo o processo ETL, bem como os resultados obtidos a partir da aplicação do DW construído.

### 4.1 ORGANIZAÇÃO, ESTRUTURA E TIPIFICAÇÃO DOS MICRODADOS ENADE 2014

Os microdados disponíveis no site são divididos nos diretórios: documentação, dados e inputs. O diretório “documentação” é composto pelo Manual do Usuário, Dicionário de Variáveis, um questionário para os estudantes de licenciaturas e outro questionário para os demais estudantes. Já o diretório “dados” possui o arquivo dos microdados do ENADE 2014 em formato CSV. Por fim, o diretório “inputs” contém os *inputs* para leitura dos microdados utilizando os softwares SAS (*Statistical Analysis System*), que consiste em um sistema integrado de aplicações para a análise de dados, e o software estatístico R.

Ao analisar o dicionário de variáveis disponibilizado, foram selecionados os dados que serão utilizados na avaliação de que elementos interferem/constroem o desempenho dos estudantes. Os dados selecionados e sua discriminação são apresentadas na Tabela 1.

Tabela 1 - Dicionário de Variáveis - ENADE 2014

| <b>NOME</b>      | <b>TIPO</b> | <b>TAMANHO</b> | <b>DESCRIÇÃO</b>   |
|------------------|-------------|----------------|--|
| nu_ano           | Numérica    | 8              | Ano de realização do exame   |
| co_orgac         | Numérica    | 8              | Código da organização acadêmica da IES   |
| co_munic_curso   | Numérica    | 8              | Código do município de funcionamento do curso  |
| co_uf_curso      | Numérica    | 8              | Código da UF de funcionamento do curso   |
| cod_regiao_curso | Numérica    | 8              | Código da região de funcionamento do curso   |
| co_curso         | Numérica    | 8              | Código do curso no ENADE   |
| nu_idade         | Numérica    | 8              | Idade do inscrito em 23/11/2014  |
| tpsexo           | Caracter    | 1              | Sexo do inscrito   |
| ano_fim_2g       | Numérica    | 8              | Ano de conclusão do 2º grau  |
| ano_in_grad      | Numérica    | 8              | Ano de início da graduação   |
| in_matutino      | Caracter    | 1              | Indicador de turno matutino  |
| in_vespertino    | Caracter    | 1              | Indicador de turno vespertino  |
| in_noturno       | Caracter    | 1              | Indicador de turno noturno   |
| tp_inscricao     | Numérica    | 8              | Indicador de concluinte / ingressante  |
| nt_obj_ce        | Numérica    | 8              | Nota bruta na parte objetiva do componente específico, convertida para escala de 0 a 100.              |
| nt_ger           | Numérica    | 8              | Nota bruta da prova – Média ponderada da formação geral (25%) e componente específico (75%) (0 a 100). |

Assim que foi concluído o processo de levantamento dos dados interessantes para a aplicação deste trabalho, foram levantadas as questões a serem analisadas a partir da utilização do DW, apresentados na próxima seção.

#### 4.2 QUESTÕES A SEREM ANALISADAS

Com o desenrolar do trabalho, através das análises realizadas e da obtenção de um conhecimento mais aprofundado sobre os microdados do ENADE, foram criados os seguintes questionamentos:

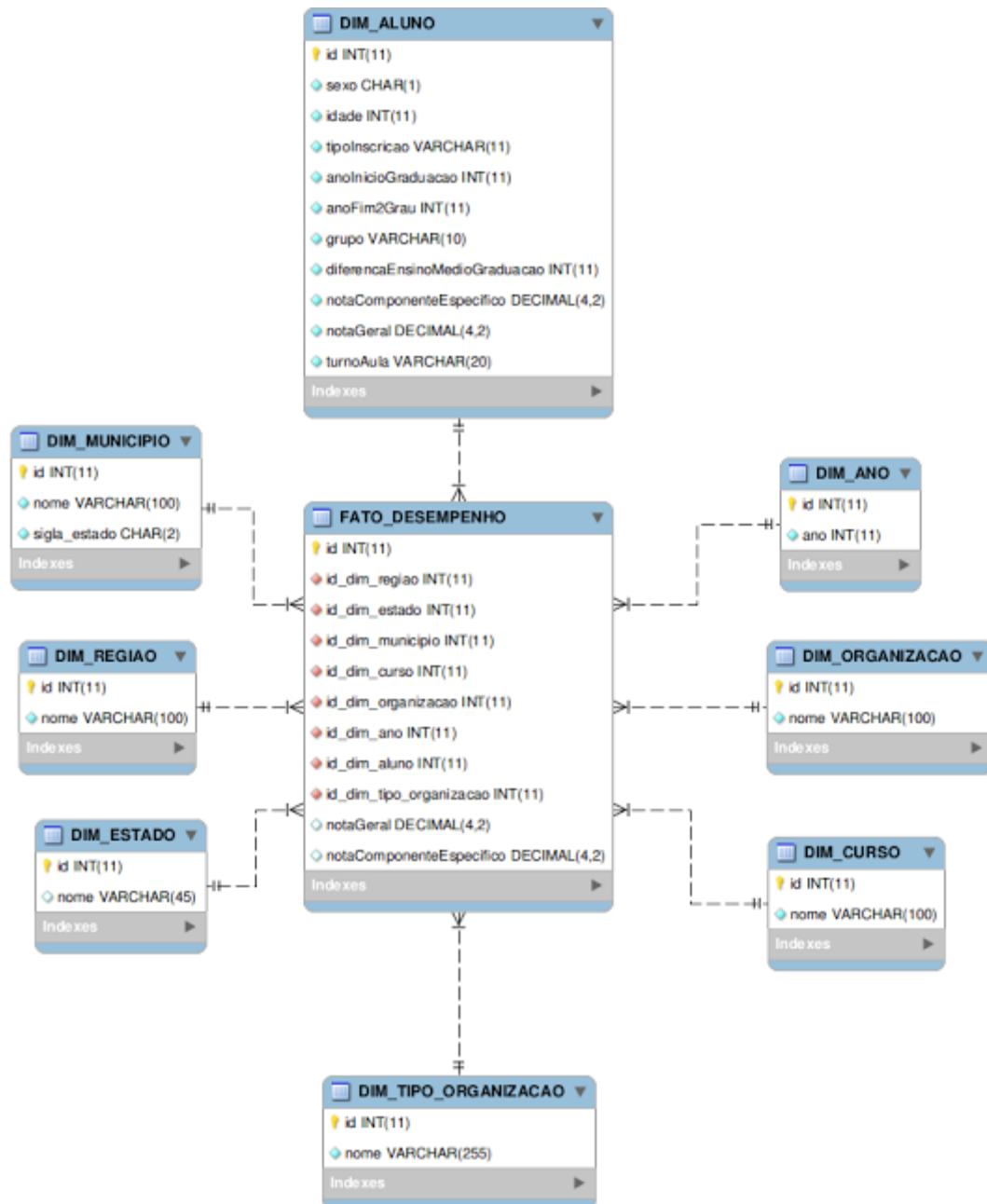
5. Qual a relação entre a idade do aluno com sua nota no ENADE com os cursos com melhores/piores desempenhos?
6. Qual a relação entre a nota no ENADE dos alunos do Estado do Tocantins com o turno do seu curso?
7. Qual o tipo de Instituição de Ensino (pública/privado) que tiveram o melhor desempenho no ENADE?
8. Qual a relação entre o período de conclusão do aluno no ensino médio e seu ingresso na faculdade com seu desempenho no ENADE?

Para resolver as questões levantadas acima, foi criado um DW, sendo apresentado na próxima seção.

#### 4.3 MODELO DE DATA WAREHOUSE CRIADO

O DW criado foi implementado através da ferramenta MySQL, utilizando a forma *top-down*, a qual exigiu um trabalho e consumo de tempo maior, como sua própria natureza impõe. A Figura 7 apresenta o DW criado.

Figura 7 - Modelagem do Data Warehouse



O modelo estrela implementado, representado na Figura 7 **Erro! Fonte de referência não encontrada.**, é composto pelas tabelas de dimensão:

- DIM\_ESTADO: corresponde aos dados dos Estados Brasileiros, sendo composta por código identificador da tabela e nome do Estado;

- DIM\_REGIAO: corresponde aos dados das Regiões Brasileiras, sendo composta por código identificador da tabela e nome da Região;
- DIM\_MUNICIPIO: corresponde aos dados dos Municípios de cada Estado Brasileiro, sendo composta por código identificador da tabela, nome do Município e sigla do Estado o qual pertence;
- DIM\_ALUNO: corresponde aos dados dos alunos avaliados, sendo composta por código identificador da tabela, nota geral na avaliação, sexo, idade, tipo de inscrição, ano de conclusão do ensino médio, ano de ingresso no curso de graduação, grupo de idade, diferença entre o ano de ingresso no curso de graduação com o ano de conclusão do ensino médio, turno do curso inscrito e nota do aluno no componente específico. Vale ressaltar que os alunos foram divididos em dois grupos: Grupo 1, alunos com até 25 anos de idade e Grupo 2, alunos com mais de 25 anos de idade. Ainda, a diferença entre o ano de ingresso do aluno no curso de graduação com o ano de conclusão do ensino médio é calculada para verificar o período de tempo que o aluno levou para ingressar em uma Instituição de Ensino Superior;
- DIM\_ANO: corresponde aos dados do ano de aplicação do exame, sendo composta por código identificador da tabela e ano de referência. Vale ressaltar que neste primeiro momento será realizada somente a carga dos dados de 2014, estando a modelagem do DW preparada para receber os dados dos próximos exames;
- DIM\_ORGANIZACAO: corresponde aos dados das Instituições de Ensino avaliadas, sendo composta por código identificador da tabela e nome da Instituição;
- DIM\_CURSO: corresponde aos dados dos cursos de graduação das Instituições de Ensino, sendo composta por código identificador da tabela e nome do curso;
- DIM\_TIPO\_ORGANIZACAO: corresponde aos dados referente ao tipo de Instituição de Ensino (pública ou privada), sendo composta por código identificador da tabela e nome do tipo de instituição.

O modelo implementado, Figura 7, representa ainda a tabela de fatos criada, sendo esta:

- FATO\_DESEMPENHO: tabela responsável por armazenar todos os atributos identificadores das tabelas de dimensões, como código identificador do estado, região, município, aluno, ano, organização, tipo de organização e curso, além da mesma possuir



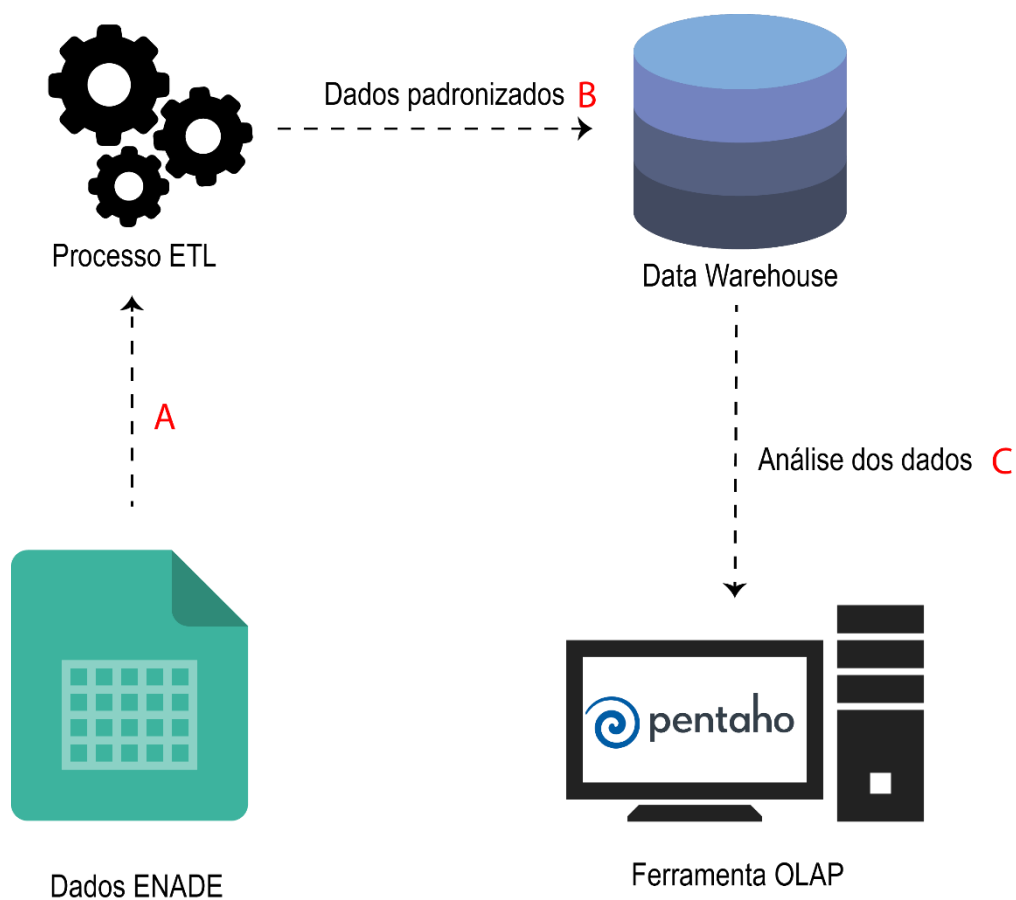
um código identificador e nota geral do aluno na avaliação e sua nota no componente específico.

A partir da criação do DW, desenvolveu-se um sistema de informação para automatizar o processo ETL, sendo apresentado na próxima seção.

#### 4.4 AUTOMATIZAÇÃO DO PROCESSO ETL

Com o objetivo de extrair, transformar e realizar a carga de dados no DW criado, desenvolveu-se um sistema de informação na linguagem de programação PHP. A Figura 8 apresenta a arquitetura do sistema implementado.

Figura 8 - Arquitetura do sistema



O fluxo do sistema iniciará a partir da extração, transformação e realização da carga de dados, sendo este processo denominado ETL, conforme pode ser observado na Figura 8-A.

Uma vez que os dados foram padronizados, realizou-se a carga dos mesmos no DW previamente criado no SGBD MySQL (Figura 8**Erro! Fonte de referência não encontrada.**-B).

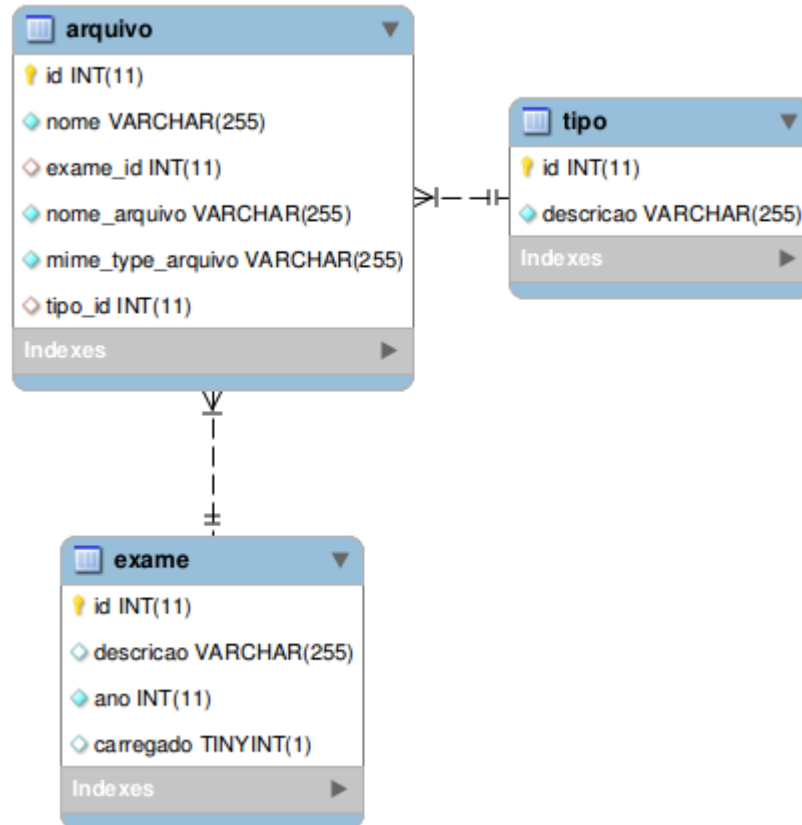
Por fim, aplicou-se a ferramenta OLAP Pentaho (Figura 8**Erro! Fonte de referência não encontrada.**-C) com o objetivo de manipular e analisar os dados em questão em diferentes perspectivas, de forma a auxiliar na descoberta de conhecimento e apoio a tomada de decisão. Esta ferramenta foi escolhida por ser *open source* e por apresentar alta confiabilidade e segurança.

A próxima seção apresenta a estrutura de banco de dados criada para armazenamento e manipulação dos dados do sistema implementado.

#### **4.4.1 Base de Dados**

O modelo relacional do banco de dados criado para abrigar e organizar as informações obtidas durante a execução do sistema implementado é apresentado na Figura 9.

Figura 9 - Modelo Relacional da Base de Dados da Aplicação



O banco de dados construído para a realização do processo ETL, representado na Figura 9, é formado pelas tabelas:

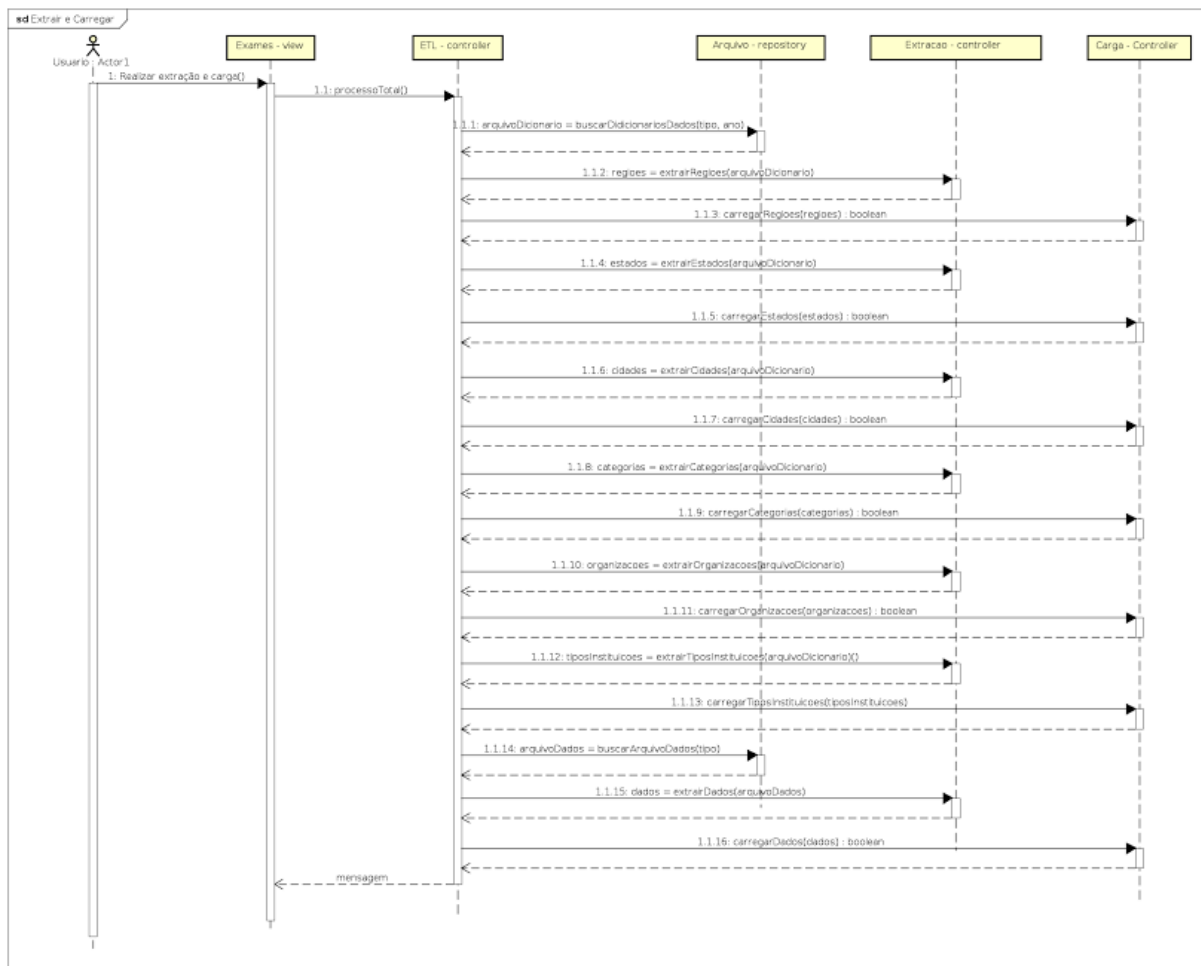
- tipo: responsável por armazenar o tipo de arquivo em questão, que variam, inicialmente, entre “dicionário” e “dados”;
- exame: consiste em registrar dados relativo ao exame, como uma nomenclatura para o exame, ano de aplicação e se este foi ou não carregado no DW;
- arquivo: tabela de relacionamento entre as demais tabelas da base de dados, responsável por armazenar um hash único identificador do arquivo (campo nome), gerado automaticamente pelo sistema; bem como o código identificador do exame, nome original do arquivo, formato do arquivo e código identificador do tipo de arquivo.

Vale ressaltar que todas as tabelas descritas possuem um atributo identificador, que foi estabelecido para um maior controle quando referido em outras tabelas. A próxima seção apresentará os métodos que compõem o sistema implementado.

#### 4.4.2 Diagrama de Sequência

Os métodos implementados para realizar a extração e carga dos dados para o DW são apresentados na Figura 10 **Erro! Fonte de referência não encontrada. Erro! Fonte de referência não encontrada.**, por meio do diagrama de sequências elaborado.

Figura 10 - Diagrama de sequência do sistema para o processo ETL.



O método destinado a realizar a extração e carga de dados é iniciado a partir da intervenção de um usuário, como apresentado anteriormente. O método “Realizar extração e carga” da View “Exames” invoca o método “processoTotal”, do controller ETL. Este, por sua vez, invoca o método “buscarDicionariosDados” do repositório “Arquivo”, sendo seu retorno utilizado como parâmetro nos métodos de extração.

O controller “Extracao” será responsável por extrair as regiões, estados, cidades, categorias (que se referem aos cursos) tipo de organização (privada, pública) e organizações (que se referem às Instituições). A cada extração, a carga do resultado obtido é realizada de forma imediata, por meio de métodos do controller “Carga”. Os métodos deste controller retornam para o controller “ETL” se a carga foi realizada ou não com sucesso, por meio de um booleano.

Concluído o processo de extração e carga, o controller “ETL” invoca o método “buscarArquivoDados”. Em seguida, invoca o método “extrairDados” do controller “Extracao”, passando o resultado obtido no método anterior como parâmetro. Além disso, o método “carregarDados” do controller “Carga” retorna o resultado obtido no último método mencionado como parâmetro, encerrando, desta forma a execução do sistema.

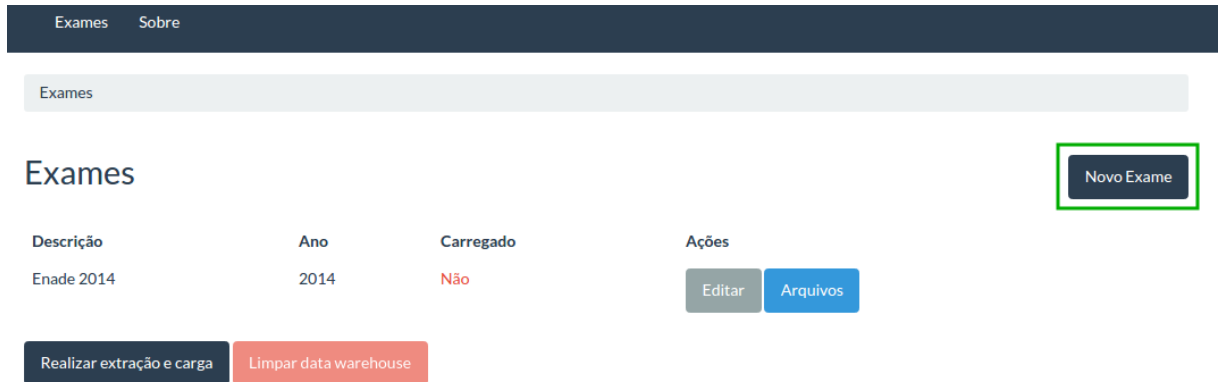
O encerramento da execução do sistema se dá por meio de uma mensagem de retorno para a view “Exames”, informando se o processo foi ou não concluído com sucesso. Caso o processo tenha sido executado com sucesso, os dados necessários para análise neste trabalho estarão gravados no DW criado para este fim.

Vale ressaltar que, no momento da carga da dimensão DIM\_ALUNO, são realizadas duas transformações. A primeira se refere ao grupo de idade que o aluno pertence. Caso ele tenha até 25 anos, o sistema insere o aluno no “Grupo 1”. Caso contrário, este é inserido no “Grupo 2”. A segunda transformação realizada está relacionada ao tempo que o aluno levou para ingressar em uma Instituição de Ensino Superior. Assim que os dados referentes ao ano de início da Graduação e de conclusão do 2º grau forem inseridos, o sistema calcula a diferença de ano(s) entre os dados mencionados, inserindo o resultado obtido no campo diferencaEnsinoMedioGraduacao.

#### **4.4.3 Telas do Sistema**

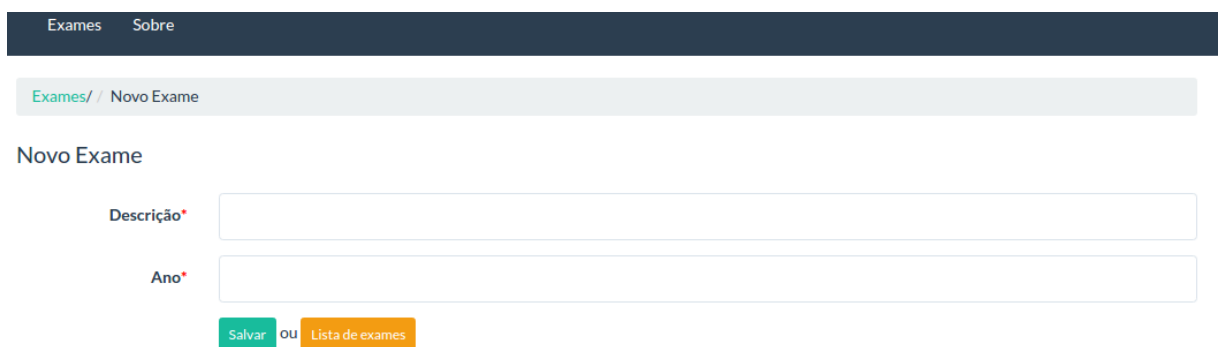
A carga dos dados do trabalho em questão será realizada diretamente no DW por meio do sistema implementado, em tempo de execução, sendo a tela inicial do sistema apresentada na Figura 11.

Figura 11 - Tela inicial do sistema



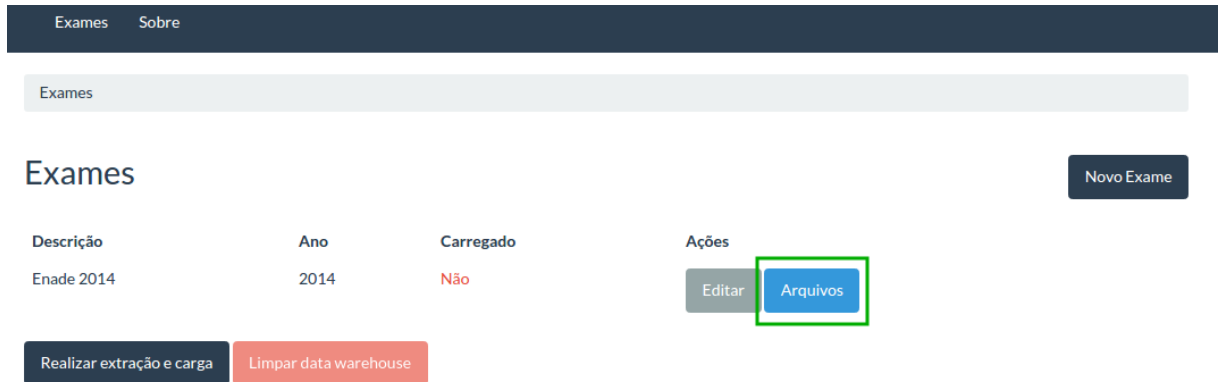
Primeiramente, o usuário deverá cadastrar um exame, por meio do botão “Novo exame”, localizado no canto superior direito da tela. Ao selecionar este botão, o usuário será direcionado para o formulário de cadastro de exame, apresentado na Figura 12. Vale ressaltar que o sistema foi implementado tendo como base o exame do ano de 2014, podendo ser utilizado nas próximas avaliações, desde que não ocorra alterações na estrutura do dicionário de variáveis.

Figura 12 - Formulário de cadastro de exame



Na tela representada na Figura 12, o usuário deverá informar um nome para o exame em questão e o ano correspondente. Ao final, o usuário deverá selecionar a opção “Salvar”, responsável por armazenar os dados informados. Vale ressaltar que será permitido apenas um cadastro de exame por ano. Concluído o cadastro, o usuário deverá informar o(s) arquivo(s) de dados, como apresentado na Figura 13.

Figura 13 - Indicando a(s) arquivo(s) de dados



A partir do cadastro do exame, o usuário poderá informar o(s) arquivo(s) de dados que deseja realizar o processo ETL, clicando no botão “Arquivos” referente ao exame desejado, como observado na Figura 13. O usuário será direcionado para a listagem de arquivos, apresentado na Figura 14.

Figura 14 - Relação de arquivos do exame



Para informar o arquivo de dados, o usuário deverá selecionar a opção “Novo Arquivo”, localizado no canto superior direito da tela representada na Figura 14. Será exibido um formulário de cadastro de arquivo, apresentado na Figura 15.

Figura 15 - Formulário de cadastro de arquivos



Exames Sobre

Exames/ / Arquivos/ / Novo Arquivo

Upload Arquivos ENADE

Tipo\* Dados

Planilha\* Escolher arquivo Nenhum arquivo selecionado

Salvar ou Lista de arquivos

No formulário de cadastro de arquivos, representado na Figura 15, o usuário deverá informar o tipo do arquivo (dicionário ou dados) e indicar o local do arquivo. Ao final, o usuário deverá selecionar a opção “Salvar”, responsável por armazenar os dados informados. Vale ressaltar que o sistema permitirá o cadastro de apenas um arquivo por tipo. Feito isso, será exibido para o usuário uma lista de arquivos do exame em questão, apresentado na Figura 16.

Figura 16 - Lista de arquivos cadastrados



Exames Sobre

Exames/ / Arquivos

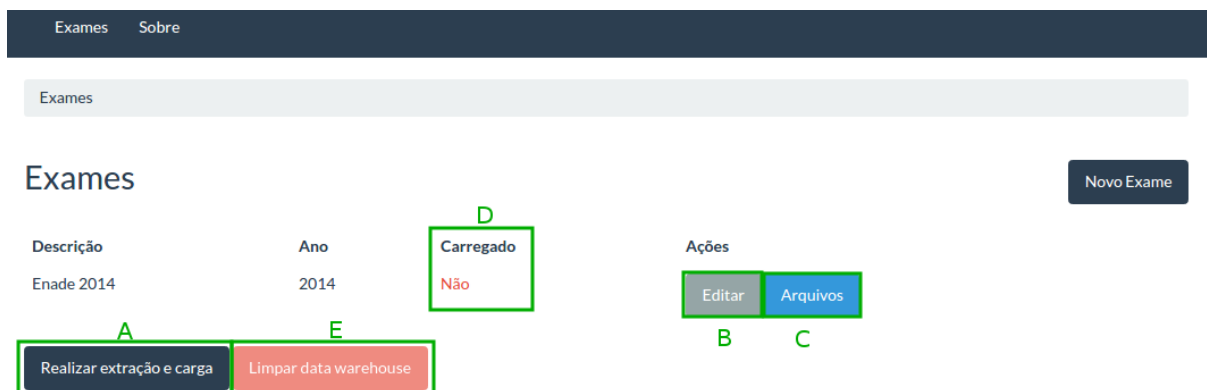
Arquivos Novo Arquivo

| Nome                        | Tipo       | Ações   |
|-----------------------------|------------|---------|
| Dicionário de Variáveis.xls | Dicionário | Deletar |
| Microdados_enade_2014.csv   | Dados      | Deletar |

Concluído o cadastro do arquivo, foram vinculados os arquivos do tipo dicionário e dados ao exame ENADE 2014, conforme Figura 16. Caso o usuário queira corrigir algum arquivo, este deverá selecionar a opção “Deletar”, localizado no registro referente ao arquivo desejado e cadastrar novamente. A Figura 17 apresenta a tela com a opção de processamento.



Figura 17 - Tela para realizar a extração e carga



Para concluir o processo, o usuário deverá selecionar a opção “Realizar extração e carga”, localizado no canto inferior esquerdo da tela representada na Figura 17**Erro! Fonte de referência não encontrada.**-A. Neste momento o sistema realizará todo o processamento necessário e, ao concluir, apresentará uma mensagem informando o sucesso ou não do processamento dos dados. Nesta tela o usuário tem ainda a opção de editar os dados do exame (Figura 17-B), bem como listar os arquivos vinculados (Figura 17**Erro! Fonte de referência não encontrada.**-C).

Cada exame cadastrado na tela representada na Figura 17 possui um “Status” (Figura 17**Erro! Fonte de referência não encontrada.**-D, tratado como uma forma de identificar se houve ou não o carregamento de dados do exame em questão no DW. A partir do momento que um exame está com o valor “Sim” no campo “Carregado” (Figura 17**Erro! Fonte de referência não encontrada.**-D), a opção de edição do exame é desabilitada, bem como o botão “Realizar extração e carga” (Figura 17**Erro! Fonte de referência não encontrada.**-A), caso todos os exames da listagem tenham sido carregados. O usuário tem a opção de “Limpar data warehouse” (Figura 17**Erro! Fonte de referência não encontrada.**-E) caso deseje apagar os dados que já foram enviados.

A partir do momento que o sistema proposto foi implementado, foram inseridos no sistema os arquivos originais de dicionário e dados do ENADE 2014, baixados diretamente no site, sendo estes carregados no DW. O arquivo com o dicionário de variáveis é utilizado para o preenchimento das tabelas de dimensões DIM\_ESTADO, DIM\_REGIAO, DIM\_MUNICIPIO, DIM\_ANO, DIM\_ORGANIZACAO e DIM\_CURSO. Já a tabela de dimensão DIM\_ALUNO

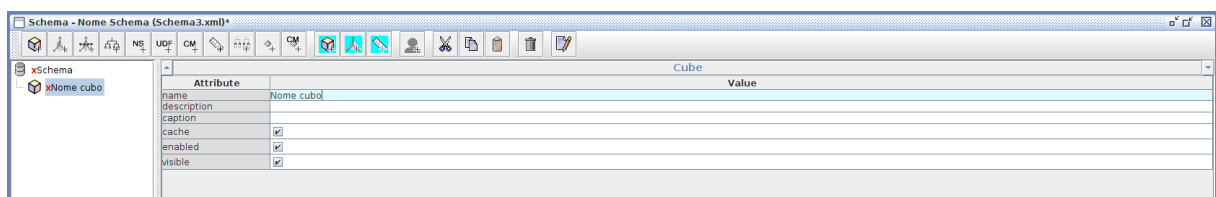
é preenchida por meio da leitura do arquivo de dados informado. Concluído o processo de carga das tabelas de dimensões, a tabela de fato FATO\_DESEMPENHO é formada pelos códigos das tabelas de dimensões e o campo nota é carregado por meio do arquivo de dados. É importante enfatizar que o próprio sistema se responsabiliza por selecionar as colunas correspondentes aos campos determinados no DW para realizar a carga de dados e desconsiderar as demais colunas que não serão utilizadas.

#### 4.5 APLICAÇÃO DO DATA WAREHOUSE PROPOSTO

Para resolver as questões levantadas na seção 4.2., foi criado um cubo de dados, ferramenta bastante utilizada na mensuração e análise de dados armazenados em um DW. Este cubo foi criado utilizando o software Modrian Schema Workbench, previamente configurado.

Afim de exemplificar a criação de um cubo no software em questão, será apresentado todo o processo realizado para criar um cubo com a tabela Região. O primeiro passo foi acessar o menu “File” da ferramenta e selecionar no submenu a opção “New”, opção “Schema”. A Figura 18 apresenta como configurar o Schema.

Figura 18 - Configurando o Schema de um Cubo

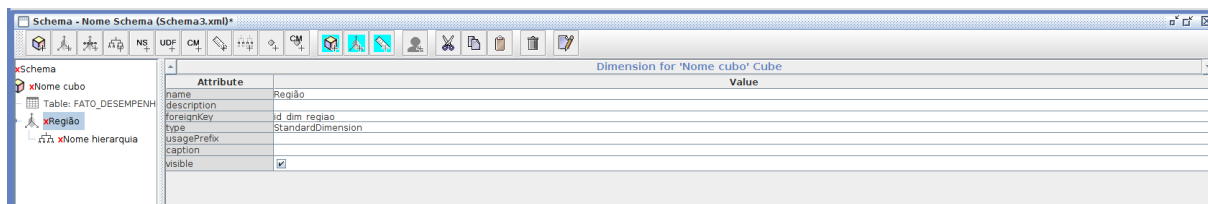


Primeiramente, para definir um Schema de um cubo, é necessário atribuir um nome para o mesmo e uma descrição, sendo a segunda atribuição não obrigatória, conforme apresentado na Figura 18. Em seguida, criou-se o cubo, clicando com o botão direito do mouse em cima da representação do Schema e, logo em seguida, selecionou-se a opção “Add cube”, atribuindo um nome para o mesmo e uma descrição, caso necessário.

Criado o cubo, faz-se necessário adicionar a tabela fato do DW. Para isso, selecionou-se o botão direito do mouse em cima da representação do Cubo criado e, em seguida, selecionou-se a opção “Add Table”, selecionando a tabela fato desejada por meio do campo “Name”. No exemplo apresentado, selecionou-se a tabela FATO\_DESEMPENHO.

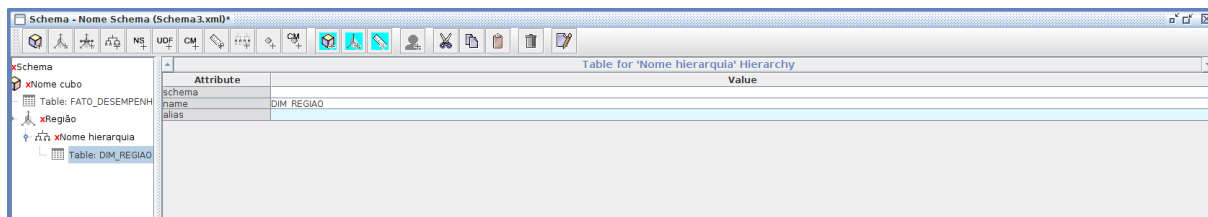
Adicionada a tabela fato no cubo, faz-se necessário adicionar a(s) tabela(s) de dimensão do DW. Para isso, selecionou-se o botão direito do mouse em cima da representação do Cubo criado e, em seguida, selecionou-se a opção “Add Dimension”, sendo necessário atribuir um nome para a mesma. Além disso, deve-se selecionar a chave estrangeira da tabela fato, conforme apresentado na Figura 19. No exemplo apresentado, atribuiu-se o nome “Região” para a dimensão e selecionou-se a chave estrangeira id\_dim\_regiao da tabela fato.

Figura 19 - Configuração da tabela de dimensão do cubo



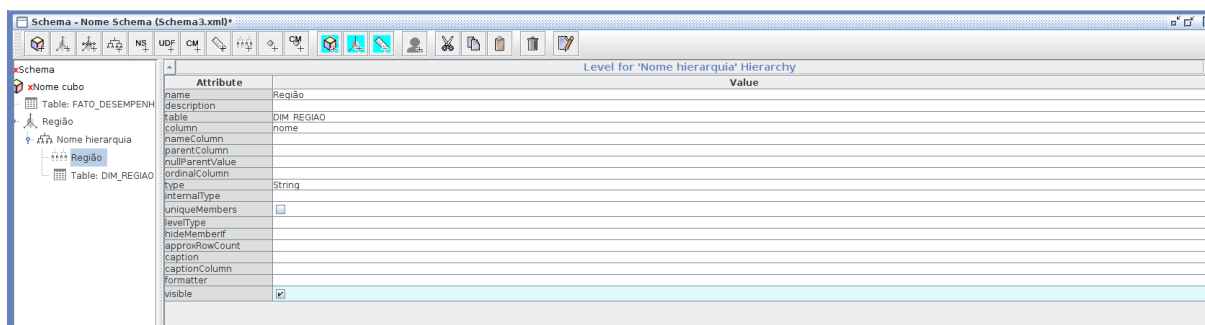
Ao expandir a tabela de fato adicionada, conforme Figura 19, é possível configurar sua hierarquia, adicionando a tabela de fato correspondente ao selecionar o botão direito do mouse em cima da representação da hierarquia, por meio da opção “Add Table”. Nesta tabela, selecionou-se a tabela de dimensão referenciada como chave estrangeira, no caso, DIM\_REGIAO, conforme apresentado na Figura 20.

Figura 20 - Configuração da tabela de hierarquia do cubo



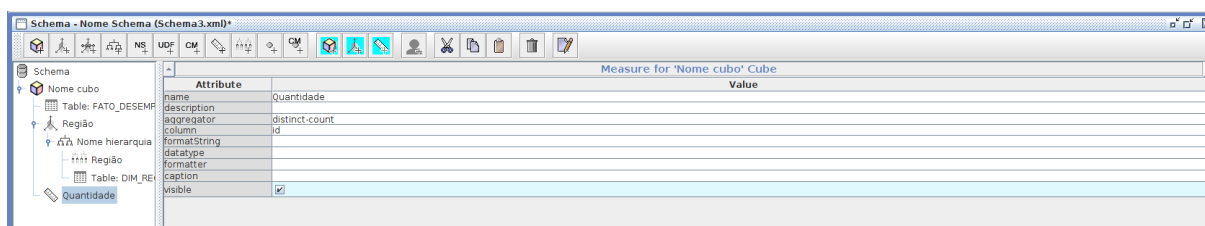
Concluída a configuração da hierarquia, adicionou-se um “Level”, ao selecionar o botão direito do mouse em cima da hierarquia e selecionar a opção “Add Level”. O “Level” se refere as propriedades da tabela que serão trabalhadas, sendo necessário atribuir um nome para o mesmo, bem como especificar a coluna desejada no campo “column” e o tipo de dado que a coluna desejada contem no campo “type”, conforme apresentado na Figura 21. No exemplo em questão, atribuiu-se o nome “Região” para o Level, selecionou-se a coluna nome da tabela DIM\_REGIAO e o tipo “String” para a coluna.

Figura 21 - Configuração do level de uma hierarquia do cubo



Uma vez adicionada uma tabela de dimensão e sua hierarquia, deve-se adicionar ao cubo uma medida a ser contabilizada, conforme apresentada na Figura 22.

Figura 22 - Configuração da propriedade quantitativa do cubo



Para configurar a propriedade quantitativa do cubo, exibida na Figura 22, selecionou-se o botão direito do mouse em cima da representação do cubo e selecionou-se a opção “Add measure”. Especificou-se um nome para a propriedade quantitativa e selecionou-se a coluna a ser contabilizada no campo “column”. No exemplo atribuiu-se o nome “Quantidade” para o “Measure” e a coluna “id” da tabela de fato FATO\_DESEMPENHO.

Como no momento em que a hierarquia foi criada não havia a especificação de nenhuma tabela, neste momento foi necessário voltar a configuração da hierarquia e selecionar o campo referente a chave primária da tabela no campo “primarykey”. Neste exemplo, selecionou-se a chave primária “id” da tabela de dimensão DIM\_REGIAO.

Por fim, para que o cubo fique disponível no software Pentaho Server, é necessário realizar a publicação do mesmo, acessando o menu “File/Publish” ainda do software Modrian Schema Workbench. Em seguida configurou-se a URL do servidor Pentaho, usuário e senha do mesmo, bem como o nome do Data Source, concluindo desta forma o procedimento de publicação.

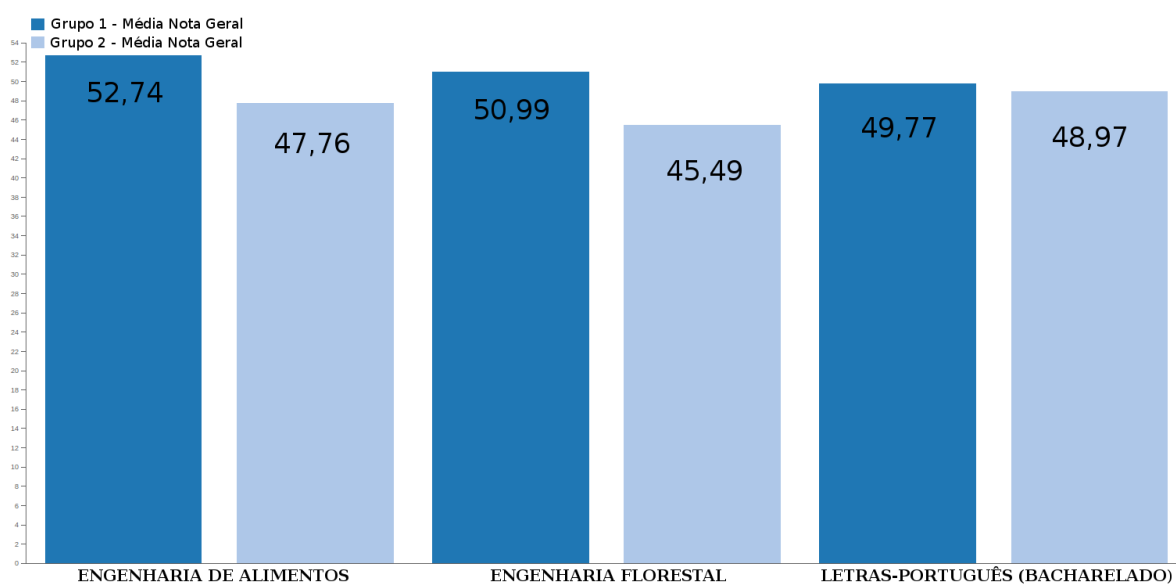
Para responder às questões levantadas neste trabalho, foram adicionadas ao cubo apresentado as demais dimensões do DW criado, apresentadas na **Erro! Fonte de referência não encontrada.**, sendo elas: Estado, Município, Aluno, Ano, Organização e Curso. Apesar

das dimensões Município e Ano não terem sido utilizadas para responder as questões levantadas, estas foram adicionadas ao cubo, deixando-o preparado para analisar novas questões. Para representar os resultados obtidos, foi utilizada o Saiku Analytics, ferramenta que oferece uma solução analítica baseada em web que permite aos usuários analisar dados corporativos, criar e compartilhar relatórios de forma rápida e fácil, apresentada na seção 3.3.3. O Apêndice C apresenta detalhes do processo de criação de gráficos a partir das informações contidas no cubo criado.

#### 4.5.1 Questão 1

A questão 1 refere-se a relação entre a idade do aluno com sua nota no ENADE com os cursos com melhores/piores desempenhos. Para responder a esta questão, foram criados dois grupos de idades. No primeiro grupo, “Grupo 1”, adicionaram-se os alunos participantes do ENADE com até 25 anos de idade, por ser a idade que normalmente um estudante está concluindo o ensino superior. Já no segundo grupo, o “Grupo 2”, adicionaram-se os alunos com mais de 25 anos.

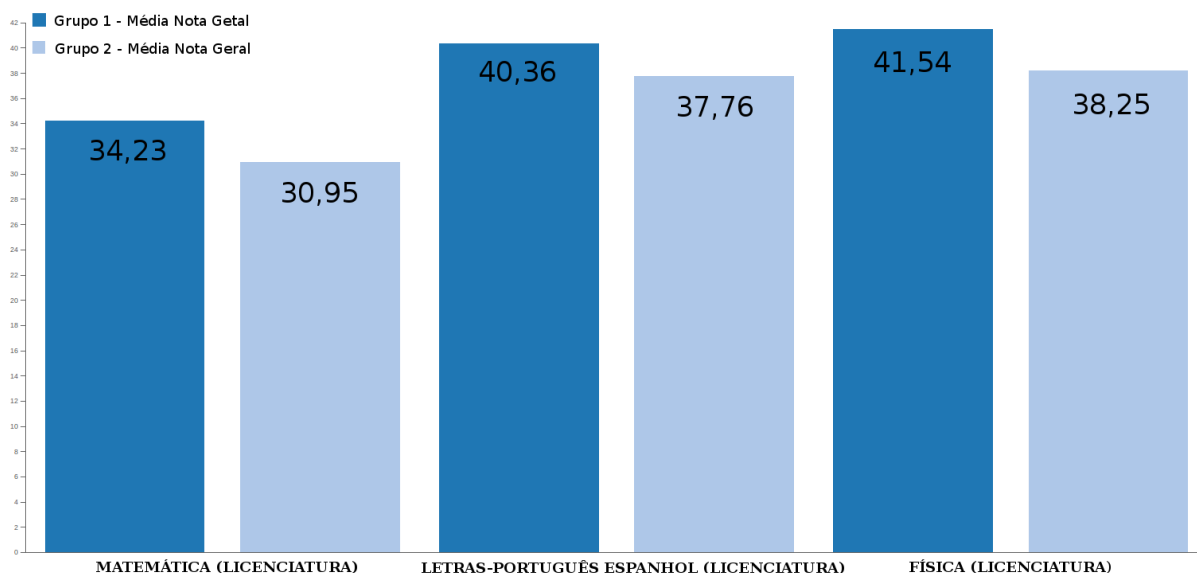
Figura 23 - Relação entre a idade do aluno com sua nota no ENADE com os cursos com melhores desempenho.



Os três cursos do país que tiveram melhor desempenho no ENADE 2014 foram Engenharia de Alimentos, Engenharia Florestal e Bacharelado em Letras-Português. No curso

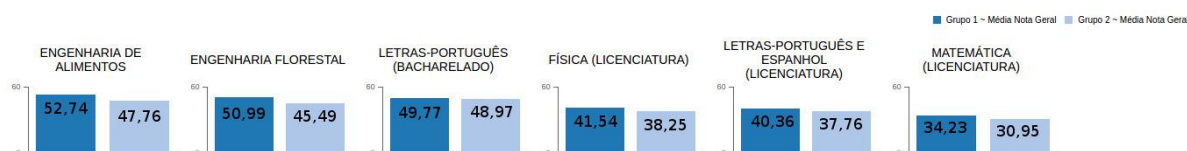
de Engenharia de Alimentos, os alunos participantes do exame inscritos no curso em questão que foram classificados no Grupo 1 fizeram 52,74 pontos e os alunos do Grupo 2 fizeram 47,76 pontos. Já para Engenharia Florestal, os alunos até 25 anos obtiveram 50,99 como média da nota e os alunos com mais de 25 anos 45,49. Por fim, os alunos participantes do curso de Licenciatura em Letras-Português que pertencem ao Grupo 1 fizeram 49,77 pontos e os do Grupo 2, 48,97 pontos. Todos os dados apresentados estão representados na **Erro! Fonte de referência não encontrada.**

Figura 24 - Relação entre a idade do aluno com sua nota no ENADE com os cursos com piores desempenho.



Os três cursos do país que apresentaram menor desempenho no ENADE 2014 foram Licenciatura em Matemática, Licenciatura em Letras-Português e Espanhol e Licenciatura em Física, respectivamente, como apresentado na **Erro! Fonte de referência não encontrada.** No curso de Licenciatura em Matemática, os alunos do Grupo 1 participantes do exame inscritos no curso em questão tiveram a média de 34,23 pontos enquanto os alunos do Grupo 2 tiveram a média de 30,95 pontos. Já para Licenciatura em Letras-Português e Espanhol, os alunos com até 25 anos de idade tiveram a média de 40,36 pontos e os alunos com mais de 25 anos 37,76 pontos. Por fim, o Grupo 1 dos alunos participantes do curso de Licenciatura em Física obteve a média de 41,54 pontos e o Grupo 2 38,25 pontos.

Figura 25 - Relação entre a idade do aluno com sua nota no ENADE com os cursos com melhores/piores desempenho.

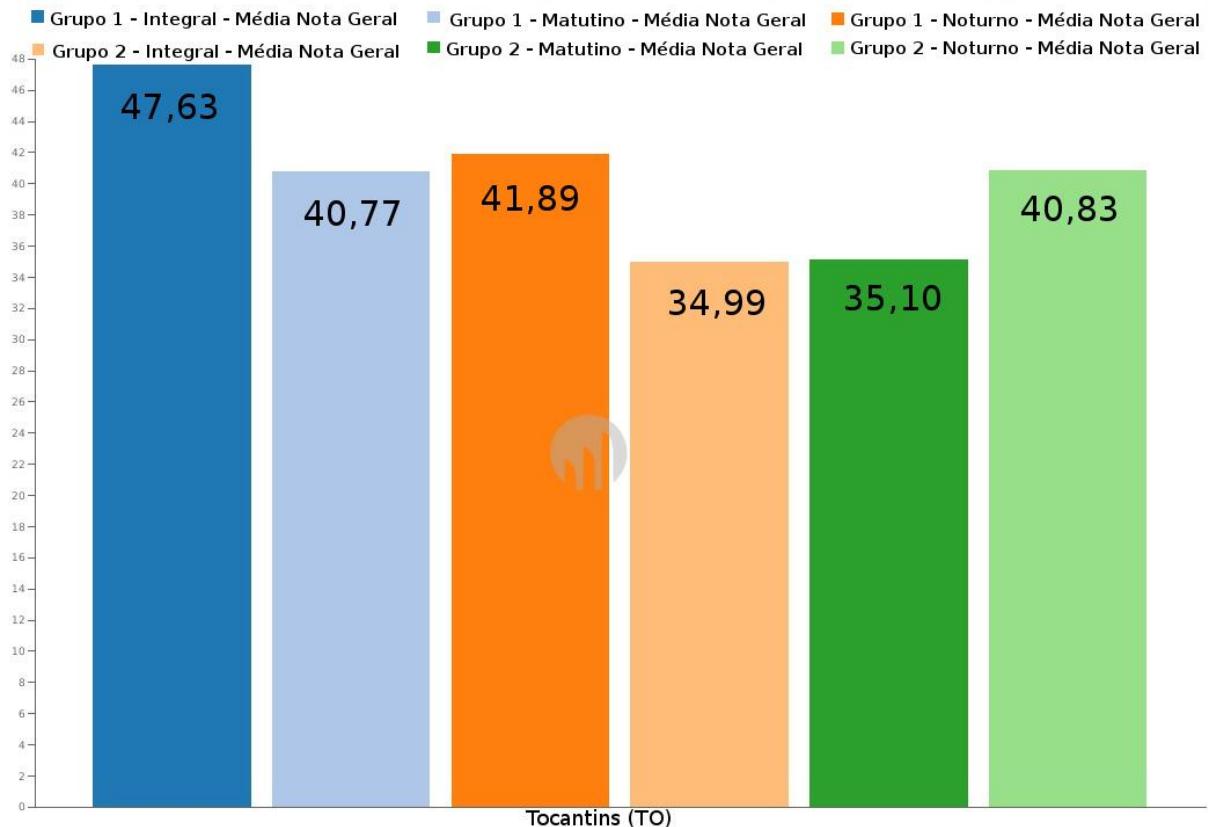


Ao realizar uma análise geral da questão 1, conclui-se que os participantes do ENADE com até 25 anos de idade obtiveram melhor desempenho no exame, independente do desempenho do curso no exame de modo geral. Sendo assim, como observado na Figura 25 **Erro! Fonte de referência não encontrada.**, os alunos mais jovens apresentaram melhores resultados tanto nos cursos melhores classificados quanto nos que receberam pior classificação. Ao contrário dos mais novos, os alunos mais velhos apresentaram baixo desempenho no exame, independente do desempenho do curso avaliado. Sendo assim, nota-se que a idade influencia diretamente no desempenho do curso, talvez pelo fato dos mais jovens terem facilidade em memorizar, na maioria das vezes aprendem com maior facilidade, bem como grande parte possui mais tempo para se dedicar aos estudos.

#### 4.5.2 Questão 2

A questão 2 refere-se a relação entre a nota no ENADE dos alunos do Estado do Tocantins com o turno do seu curso. Para responder a esta questão, foi adotada o mesmo agrupamento utilizado na questão 1, apresentado na seção 4.5.1, onde o “Grupo 1” corresponde aos alunos participantes do ENADE com até 25 anos de idade e o “Grupo 2” aos alunos com mais de 25 anos. Para cada grupo foi levantada a média dos alunos por turno, que variam em Integral, Matutino e Noturno.

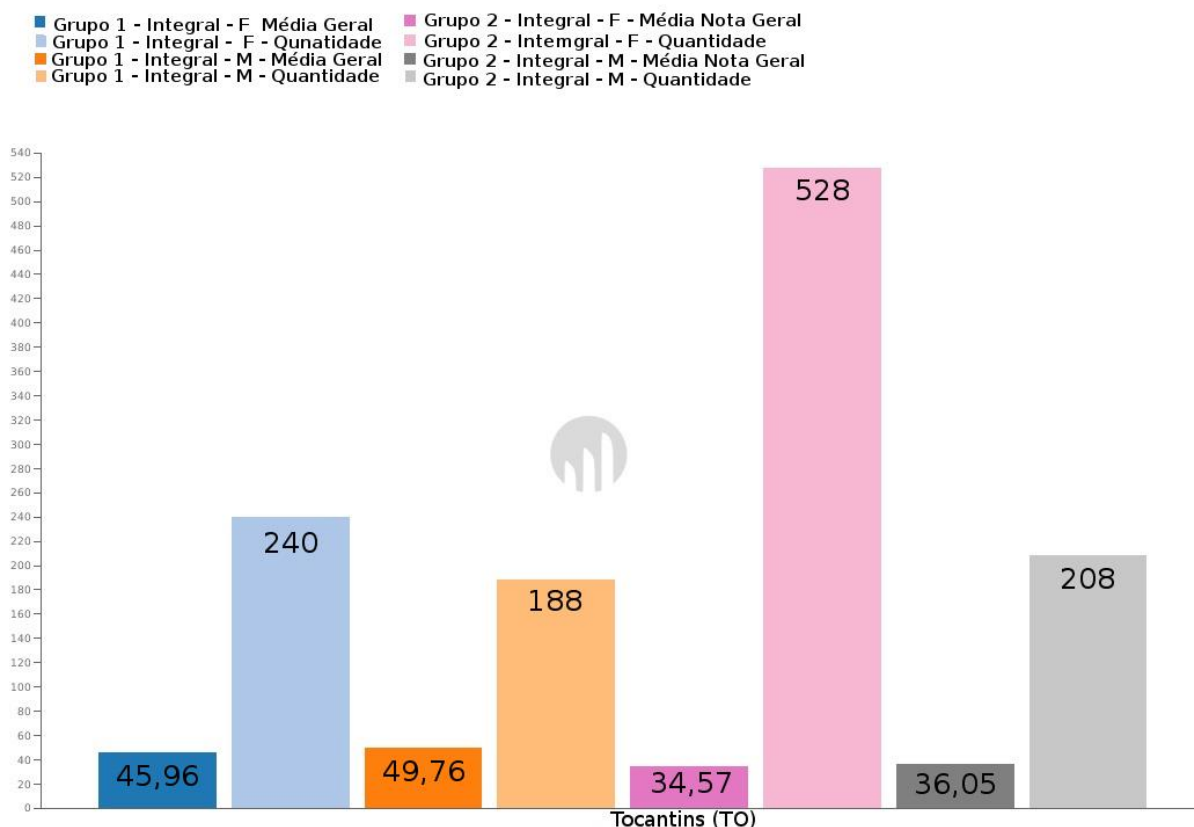
Figura 26 - Relação entre a nota no ENADE dos alunos do Estado do Tocantins com o turno do curso que cursam.



Ao analisar o Grupo 1, observa-se que os alunos que estudam em período integral apresentaram melhor desempenho no ENADE (47,63 pontos), seguidos dos alunos do turno noturno (41,89 pontos) e matutino (40,77 pontos), respectivamente (**Erro! Fonte de referência não encontrada.**). Já no Grupo 2, os alunos do turno noturno apresentaram melhor desempenho no exame (40,83 pontos), seguidos pelos alunos do turno matutino (35,10 pontos) e integral (34,99 pontos), respectivamente (**Erro! Fonte de referência não encontrada.**). Destaca-se que o desempenho dos alunos mais jovens em cursos de período integral é bem superior se comparado aos demais turnos. Já os alunos mais velhos apresentam melhor desempenho nos cursos noturnos. Na tentativa de identificar uma característica e/ou fator que influencie nos valores obtidos nos cursos de período integral, subdividiu-se os grupos de pessoas por sexo, conforme apresentado na **Erro! Fonte de referência não encontrada.**



Figura 27 – Relação entre a nota no ENADE dos alunos do Estado do Tocantins dos cursos com período integral com sua idade e sexo.



Ao analisar os alunos inscritos em cursos de período integral e subdividindo-os em de grupos de idade (Grupo 1: até 25 anos, e Grupo 2: acima de 25 anos) e sexo (feminino e masculino), observa-se que as 240 mulheres do Grupo 1 que fizeram o exame tiveram a média de 45,96 pontos, seguidas por 188 homens com a mesma faixa etária que tiveram a média de 49,76 pontos. Apesar da quantidade de mulheres deste grupo ser maior que a quantidade de homens, os homens obtiveram uma maior pontuação no exame aplicado (**Erro! Fonte de referência não encontrada.**). Com relação ao Grupo 2, 528 inscritos são do sexo feminino e 208 são do sexo masculino. Assim como no Grupo 1, apesar da quantidade de inscritos do sexo feminino ser superior a quantidade de inscritos do sexo masculino, os homens obtiveram um melhor desempenho na nota do exame, com 36,05 pontos contra 34,57 pontos das mulheres inscritas.

Ao realizar uma analogia na **Erro! Fonte de referência não encontrada.**, nota-se que as mulheres de ambos os grupos tiveram um menor desempenho no exame, mesmo representando a maior ocorrência no que se refere ao quantitativo de alunos. Contudo, as

mulheres mais jovens apresentaram melhor desempenho no exame, por provavelmente não serem casadas e não terem filhos. Já as mulheres mais velhas provavelmente trabalham fora, são casadas e possuem filhos, sendo desta forma mais complicado conciliar seus horários com os de um curso de tempo integral.

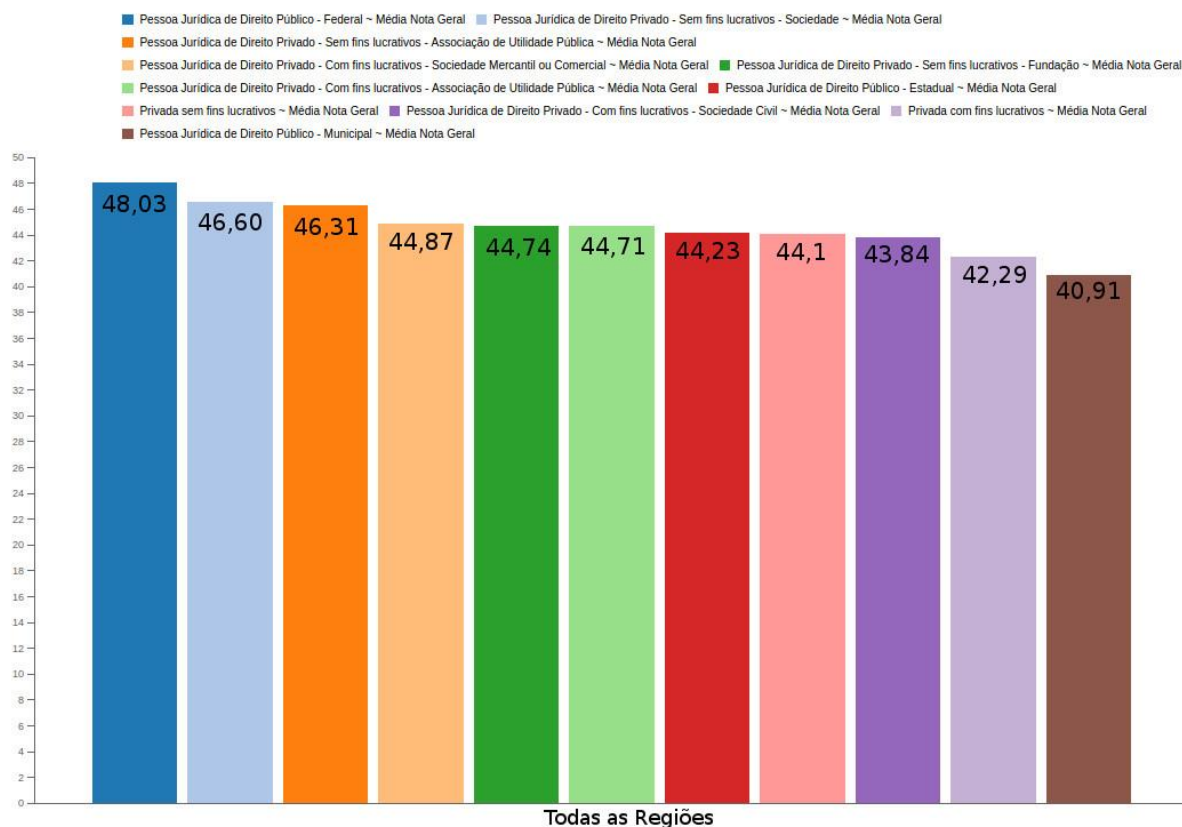
A mesma analogia pode ser realizada com os alunos do sexo masculino, sendo que os mais velhos apresentam menor desempenho nos cursos de período integral. Sugere-se que o fato destes possuírem um emprego e responsabilidades familiares afeta diretamente em seu desempenho acadêmico. Para os homens, ao observar a **Erro! Fonte de referência não encontrada.**, nota-se que o melhor horário para os estudos seria o período noturno.

De forma geral, ao analisar os dados obtidos, observou-se que o sexo não é um fator determinante para o desempenho dos alunos avaliados, embora tenham apresentado uma pequena diferença nos resultados entre os grupos.

#### 4.5.3 Questão 3

A questão 3 está relacionada ao tipo de Instituição de Ensino que obteve melhor desempenho no ENADE 2014. Além do tipo, se públicas ou privadas, as instituições estão divididas também em esferas (federal, estadual e municipal), como apresentado na **Erro! Fonte de referência não encontrada.**

Figura 28 - Relação entre o tipo de Instituição de Ensino (pública/privado) que tiveram o melhor desempenho no ENADE



Como pode ser observado na **Erro! Fonte de referência não encontrada.**, as instituições Federais obtiveram o melhor desempenho no exame, com a média de 48,03 pontos, seguidas pelas instituições privadas sem fins lucrativos – sociedade e associação de utilidade pública, com 46,60 e 46,31 pontos, respectivamente. Em seguida encontram-se as instituições privadas com fins lucrativos – sociedade mercantil ou comercial com 44,87. Ainda, instituições sem fins lucrativos – fundação, com 44,74 e instituições com fins lucrativos – associação de utilidade pública com 44,71. Têm-se também as instituições estaduais e sem fins lucrativos, com 44,23 e 44,1 pontos, respectivamente. Instituições com fins lucrativos – sociedade social – e com fins lucrativos de modo geral, obtiveram a média de 43,84 e 42,29 pontos concomitantemente. Por fim, as instituições municipais apresentaram menor desempenho com a média de 40,91 pontos.

Ao analisar o gráfico em questão, observa-se que a diferença do desempenho de uma instituição para a outra decresce com uma variação mínima no valor de suas médias gerais.

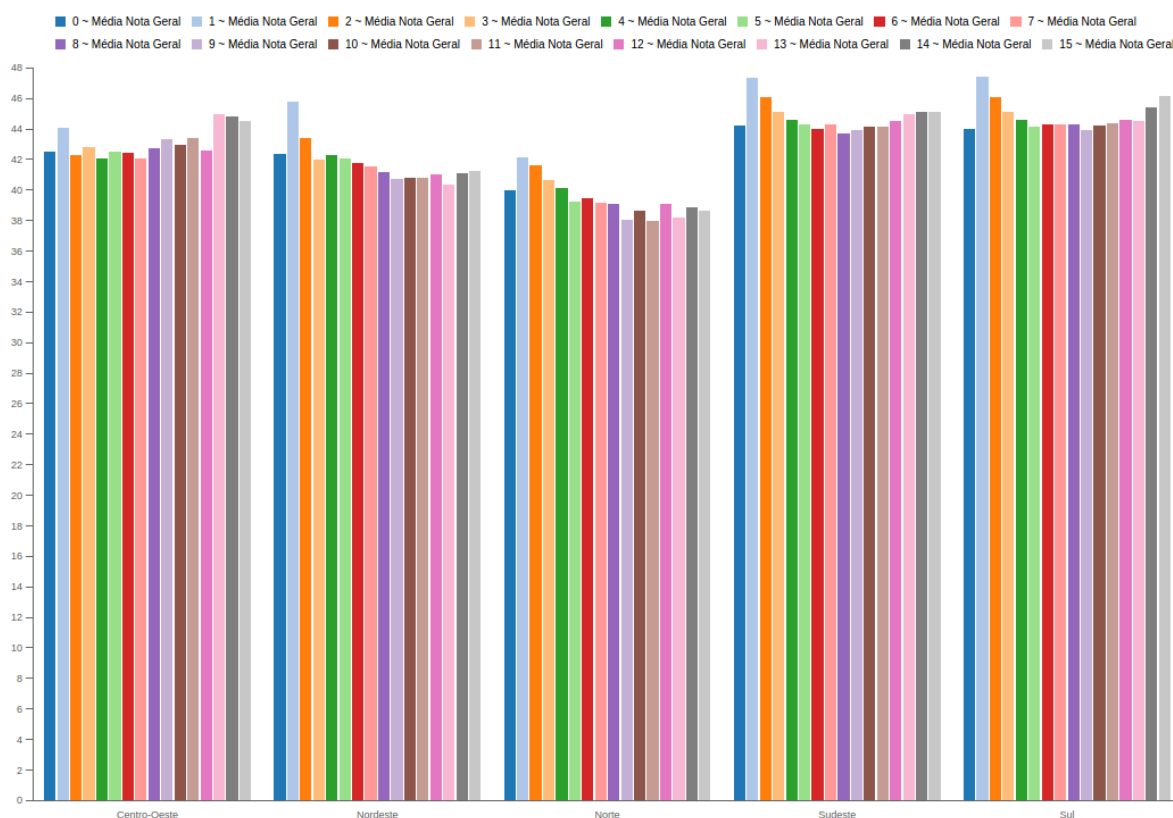
Classificando as instituições apenas pelos tipos privada e pública, identificaram-se apenas três instituições públicas e as demais privadas, totalizando oito instituições. As instituições públicas podem ser classificadas nas esferas federal, estadual e municipal. Dentre estas, as instituições federais apresentaram maior desempenho, seguidas das estaduais e municipais, respectivamente. Uma hipótese levantada para justificar o resultado apresentado é o fato das verbas federais serem maiores que as demais esferas, bem como o fato que na maioria das vezes estas instituições exigem mais dos candidatos no processo seletivo do que as instituições das esferas estadual e municipal, sendo mais criteriosos.

No que se diz respeito às instituições privadas, estas podem ser classificadas em: com fins lucrativos e sem fins lucrativos, havendo quatro instituições de cada uma das duas classificações. Dentre as instituições com fins lucrativos, encontram-se a de Sociedade Mercantil ou Comercial, Associação de Utilidade Pública, Sociedade Civil e as demais instituições privadas que não se encaixam nestas classificações. Já entre as instituições sem fins lucrativos, encontram-se a de Sociedade, Associação de Utilidade Pública, Fundação e as demais instituições privadas que não se encaixam nas classificações apresentadas. De forma geral, ao realizar um comparativo entre as instituições com fins lucrativos com as sem fins lucrativos, percebe-se que as instituições sem fins lucrativos foram melhor classificadas quanto ao desempenho obtido no ENADE. Julga-se que este resultado foi obtido devido ao fato das instituições em questão focarem mais na qualidade do aprendizado de seus alunos para manter sua característica econômica.

#### **4.5.4 Questão 4**

A questão 4 está relacionada ao período de conclusão do aluno no ensino médio e seu ingresso na faculdade com seu desempenho no ENADE 2014. Destaca-se que para esta questão foram analisados os alunos que levaram até 15 anos para ingressar em uma instituição de ensino superior desde que concluíram o ensino médio. A Figura 29 representa a análise gráfica desses dados.

Figura 29 - Relação entre o período de conclusão do aluno no ensino médio e seu ingresso na faculdade com seu desempenho no ENADE



Os alunos da região Centro-Oeste que levaram de 13 a 15 anos para ingressar em uma instituição de ensino superior apresentaram melhor desempenho no exame. Já os alunos que levaram 2, 4 e 7 anos para ingressar neste tipo de instituição receberam as menores médias na avaliação. No Nordeste, os alunos que levaram apenas 1 ano para ingressar em uma instituição de ensino superior apresentaram melhor desempenho no exame, apresentando uma média bem acima que os demais grupos. Entre os que receberam as menores médias da região na avaliação, destacam-se os que levaram 13, 9 a 11 anos para ingressar em uma instituição superior, havendo uma mínima variação entre os grupos de 9 a 11 anos. Já os alunos da região Norte que levaram de 1 a 3 anos para ingressar em uma instituição de ensino superior apresentaram melhor desempenho no ENADE. Por outro lado, os alunos que levaram 9, 11 e 13 anos para ingressar neste tipo de instituição receberam as menores médias no exame. No Sudeste, os alunos que levaram de 1 a 3 anos para ingressar em uma instituição de ensino superior apresentaram melhor desempenho no exame. Entre os que receberam as menores médias da região na avaliação, destacam-se os que levaram 8, 9 e 6 anos para ingressar em uma instituição superior. Contudo,

a variação das melhores e piores médias da região é mínima, tendendo para uma média alta. Por fim, os alunos da região Sul que levaram de 1, 2 e 15 anos para ingressar em uma instituição de ensino superior apresentaram melhor desempenho no ENADE. Por outro lado, os alunos que no mesmo ano de conclusão do ensino médio e os que levaram 9, 5 e 10 anos para ingressar neste tipo de instituição receberam as menores médias no exame. Porém, assim como pode ser observado na região Sudeste, a variação das melhores e piores médias da região é mínima, tendendo para uma média alta.

Ao analisar as regiões de uma forma geral, as regiões Sul e Sudeste obtiveram as melhores médias do ponto de vista do período que seus alunos levaram para ingressar em uma instituição de ensino superior desde que concluíram o ensino médio. Já a região Norte, seguida das regiões Nordeste e Centro-Oeste, respectivamente, obtiveram as médias mais baixas no quesito em questão. Sendo assim, observa-se que o tempo que o aluno levou para ingressar em uma instituição de nível superior teve uma menor influência no desempenho das regiões Sudeste e Sul no ENADE quando comparadas às demais regiões, especialmente quando comparada à região Norte.

Outro ponto a ser observado é o comportamento do gráfico de forma geral. A medida que os alunos levam mais tempo para ingressar em uma instituição de nível superior, sua média geral de desempenho na avaliação decresce até certo ponto. Deste ponto em diante, há novamente um crescimento na média da região, ocorrendo aproximadamente no mesmo período em grande parte delas. A partir desta análise, acredita-se que seja interessante realizar um estudo para identificar que fatores influenciaram a decadência e crescimento no desempenho das regiões, bem como realizar um levantamento de características comuns dos alunos de cada grupo.

#### **4.5.5 Considerações**

Como apresentado nas seções 4.5.1 a 4.5.2, que se referem as questões analisadas neste trabalho, conclui-se que os cursos de Engenharia de Alimentos, Engenharia Florestal e Bacharelado em Letras-Português são os cursos de nível superior melhor classificados no ENADE 2014. Já os cursos de Licenciatura em Física, Letras-Português e Espanhol e Matemática tiveram as piores classificações na avaliação.

No estado do Tocantins, os cursos com turno integral foram melhor classificados no exame do ENADE, estando matriculados mais mulheres do que homens. Já os cursos do período matutino receberam as menores médias na avaliação.

Quanto aos tipos de instituições de ensino, as da esfera federal foram melhor classificadas no ENADE 2014, seguidas das instituições privadas sem fins lucrativos. Em seguida aparecem as instituições estaduais, bem como as com fins lucrativos e municipais.

Por fim, no que se refere ao período que os participantes levam para ingressar em uma instituição de ensino superior desde que concluem o ensino médio, observou-se no gráfico apresentado que a medida que os alunos levam mais tempo para ingressar em uma instituição de nível superior, sua média geral de desempenho na avaliação decresce até certo ponto. De um ponto em diante, há novamente um crescimento na média da região, ocorrendo aproximadamente no mesmo período em grande parte delas.

## 5. CONSIDERAÇÕES FINAIS

Ao elaborar este trabalho, teve-se como um dos focos a realização de pesquisas e estudos referentes à Data Warehouse de forma a compreender seus principais conceitos, características, modelos e formas de implementação para uma posterior aplicação, aqui também apresentada. A ideia do trabalho em questão consiste na construção de um Data Warehouse (DW), no qual serão armazenados dados extraídos da base do ENADE 2014. Este DW possibilitará produzir dados referenciais que permitam a definição de ações voltadas à melhoria da qualidade dos cursos de graduação das Universidades Brasileiras. Esta melhoria se dará por meio da aplicação de uma ferramenta OLAP sobre os dados armazenados de tal forma a possibilitar uma melhor análise dos dados do ENADE 2014, segundo foco deste trabalho.

Com a implementação do ambiente de DW apresentado, pode-se constatar que tanto a tomada de decisão quanto a busca por informações é realizada de maneira ágil. Além disso, alguns dados foram transformados, gerando novas informações relevantes para análise, como ocorreu com o grupo de faixa etária criado, bem como a informação sobre o período que pessoas levaram para entrar em uma Instituição de Ensino de nível Superior desde que saíram do Ensino Médio.

A implementação de um sistema para automatizar o processo de extração, transformação e carga de dados no DW foi essencial para agilizar e facilitar este procedimento, que corresponde a primeira etapa envolvida no método de aquisição de conhecimento.

O sistema apresentado na seção 4.4 pode ser manuseado por qualquer pessoa, independentemente de sua área de atuação. Além disso, este já está preparado para receber os dados de futuros exames. Contudo, vale ressaltar que caso o INEP altere algum ponto na estrutura dos microdados do ENADE será necessário realizar adequações nos produtos obtidos neste trabalho, sendo eles: DW e sistema responsável pelo processo ETL.

Durante a análise das questões levantadas, apresentadas na seção 4.5, foi identificado que:

- pessoas mais jovens possuem as melhores médias na avaliação quando comparadas aos mais velhos;
- os cursos de graduação apresentaram melhor desempenho no exame quando comparados às Licenciaturas;



- os alunos mais jovens apresentam maior rendimento em cursos de tempo integral. Já alunos mais velhos foram melhores classificados em cursos noturnos;
- há mais mulheres do que homens cursando instituições de ensino superior;
- os homens, independentemente de sua idade, apresentam um melhor aproveitamento em cursos noturnos;
- as instituições federais apresentaram melhor desempenho na avaliação do ENADE, enquanto as municipais apresentam um desempenho mais baixo;
- as instituições sem fins lucrativos apresentaram uma melhor classificação quando comparadas as instituições com fins lucrativos;
- as regiões Sul e Sudeste obtiveram as melhores médias do ponto de vista do período que seus alunos levaram para ingressar em uma instituição de ensino superior desde que concluíram o ensino médio enquanto a região Norte obteve médias baixas.

A partir dos resultados obtidos, foram identificados fatores que influenciam diretamente na classificação do curso e/ou instituição, sendo eles:

- a idade do candidato;
- o sexo do candidato;
- o fato do candidato possuir um emprego ou não;
- o fato do candidato ser casado;
- o fato do candidato ter filhos;
- os investimentos financeiros aplicados às Instituições;
- a qualidade e cobrança do processo seletivo de ingresso à Instituição de ensino superior;
- o ano de conclusão do candidato no ensino médio e seu ingresso em um curso superior.

De forma geral, este trabalho produziu três produtos: um Data Warehouse, um sistema de automatização do processo ETL e análises com questões que podem influenciar o desempenho dos alunos em exames realizados para avaliar a qualidade de um curso superior e/ou Instituição.

Como proposta para trabalhos futuros, pretende-se fazer a carga de dados dos próximos exames a serem aplicados e analisar quais os resultados obtidos a cada nova aplicação, bem

como realizar um comparativo entre os anos em questão, de forma a verificar a evolução ou regressão dos pontos levantados neste trabalho. Além disso, pretende-se formular novas questões a serem analisadas, a fim de identificar outros fatores que afetam o desempenho de uma instituição na avaliação do ENADE. Como exemplo, de que forma o investimento aplicado as Instituições Públicas está ou não interferindo em seu desempenho de modo geral. Acoplar o Exame Nacional de Ensino Médio (ENEM) à ferramenta apresentada também é uma pretensão futura, com o intuito de identificar se o aluno já vem apresentando baixo desempenho desde o ensino médio. Por fim, pretende-se adaptar o sistema ETL implementado, de forma que o usuário possa configurar por meio de uma tela do sistema os dados necessários de conexão com o DW desejado.

## 6. REFERÊNCIAS

ANZANELLO, Cynthia Aurora. **OLAP Conceitos e Utilização**. 2002. 7f. Artigo – Universidade do Rio Grande do Sul – Porto Alegre. Disponível em: <<http://www.softsystemitead.com.br/phocadownload/BI/Conceitos%20OLAP.pdf>>. Acessado em: 24 ago. 2016.

BATISTA, L. P. L. **Implementação de uma solução open source de Business Intelligence com metodologia Ágil**. 2014. 73 f. Dissertação (Mestrado em Gestão de Informação) – Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Lisboa, 2014. Disponível em: <<https://run.unl.pt/bitstream/10362/13763/1/TGI0017.pdf>>. Acessado em: jun. 2017.

BOUMAN, Roland; DONGEN, Jos Van. **Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL**. Indianápolis: Wiley Publishing, 2009. 604 p.

BUYENS, J. **Aprendendo MySQL e PHP**. 1. ed. São Paulo: Makron Books, 2002.

Cognos IBM. Publicação Eletrônica em IBM. Disponível em: <<http://www-03.ibm.com/marketing/br/campanhafpm/>>. Acessado em: 23 ago. 2016.

Data Warehouse – Data Mining. Disponível em: <<http://www.datawarehousing.inf.br/>>. Acessado em: 3 set. 2011.

DBMiner E1.1. **User Manual – For Windows NT/95**. DBMiner Technology Inc. Disponível em: <<http://db.cs.sfu.ca/DBMiner/download2>>. Acessado em: 3 set. 2011.

FERNEDA, E. **Recuperação de Informação: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. 2003. 147 f. Tese (Doutorado em Ciência da Comunicação) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2003. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/pt-br.php>>. Acessado em: 24 ago. 2016.

GOLDWIN. Pentaho Open Source BI. Disponível em: <<http://bcinews.files.wordpress.com/2010/02/pentaho-open-source-bi.ppt>>. Acessado em: 23 ago. 2016.

HOKAMA, Daniele Del Bianco et. al. **A Modelagem de Dados no Ambiente Data Warehouse**. Universidade Presbiteriana Mackenzie, Faculdade de Computação e Informática, São Paulo, 2004. Disponível em: <<http://meusite.mackenzie.com.br/rogerio/tgi/2004ModelagemDW.pdf>>. Acessado em: 24 ago. 2016.

INEP. Exame Nacional de Desempenho dos Estudantes ENADE. 2011. Disponível em: <<http://portal.inep.gov.br/enade>>. Acessado em: 25 ago. 2016

INMON, W. H. **Como construir o Data Warehouse**. 2ª ed. New York: Editora Campus, 1997.

KIMBALL, Ralph et. al. **The data warehouse lifecycle toolkit: expert methods for designing, developing and developing data warehouse**. New York: John Wiley & Sons, 1998. Disponível em: <<http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471255475.html>>. Acessado em: 3 set. 2016.

MACHADO, F. N. R.; ABREU, M. P. **Projeto de Banco de Dados: uma visão prática**. São Paulo: Érica, 1996.

MAZZOLA, Irany Salgado. 2002. 153p. **Projeto de Data Warehouse Dimensional**. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis, 2002. Disponível em: <<https://repositorio.ufsc.br/bitstream/handle/123456789/83465/184713.pdf?sequence=1>>. Acessado em: 3 set. 2016.

Microstrategy. Publicação Eletrônica em Microstrategy. Disponível em: <<http://www.microstrategy.com.br/Software/businessintelligence>>. Acessado em: 23 ago. 2016

MOOERS, C. Zetocoding applied to mechanical organization of knowledge. **American Documentation**, v. 2, n. 1, p. 20-32. Disponível em: <<https://courses.engr.illinois.edu/cs473/fa2013/misc/zetocoding.pdf>>. Acessado em: 23 ago. 2016.

NIEDERAUER, Juliano; PRATES, Rubens. **Guia de Consulta Rápida MySQL 5**. Novatec Editora Ltda. 2006. Disponível em: <<http://www.martinsfontespaulista.com.br/anexos/produtos/capitulos/209813.pdf>>. Acessado em: 24 ago. 2016.

OGGIO, P. D. PHP: Programando com Orientação a Objetos. Publication Place: NOVATEC, 2015. Disponível em: <[https://books.google.com.br/books?hl=en&lr=&id=yEPOCgAAQBAJ&oi=fnd&pg=PT4&dq=php+conceito&ots=n1Dlit8vW3&sig=W\\_9j6yUpDC4x2iqxoIpqIHkuxW4#v=onepage&q=hp%20conceito&f=false](https://books.google.com.br/books?hl=en&lr=&id=yEPOCgAAQBAJ&oi=fnd&pg=PT4&dq=php+conceito&ots=n1Dlit8vW3&sig=W_9j6yUpDC4x2iqxoIpqIHkuxW4#v=onepage&q=hp%20conceito&f=false)>. Acessado em: 15 mai. 2017.

PENTAHO, Open Source Business Intelligence. **Getting Started With Pentaho Data Integration**. 2010.

PEREIRA, Walter A. L. **Uma metodologia de inserção de tecnologia de data warehouse em organizações.** 2000. Trabalho Individual II (Mestrado em Ciência da Computação) – Instituto de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

POE, Vidette; KLAUER, Patricia; BROBST, Stephen. **Building a data warehouse for decision support.**New Jersey, Prentice Hall PTR. 1998.

PORTIFÓLIO SAP BUSINESSOBJECTS. Disponível em:  
<<http://www.brazil.businessobjects.com/>>. Acessado em: 23 ago. 2016.

REBOUÇAS, F. **Data Warehouse.** Disponível em:  
<<http://www.infoescola.com/informatica/data-warehouse/>>. Acesso em: 24 ago. 2016.

SAS. Publicação em SAS. Disponível em: <<http://www.sas.com/software/>>. Acessado em: 23 ago. 2016.

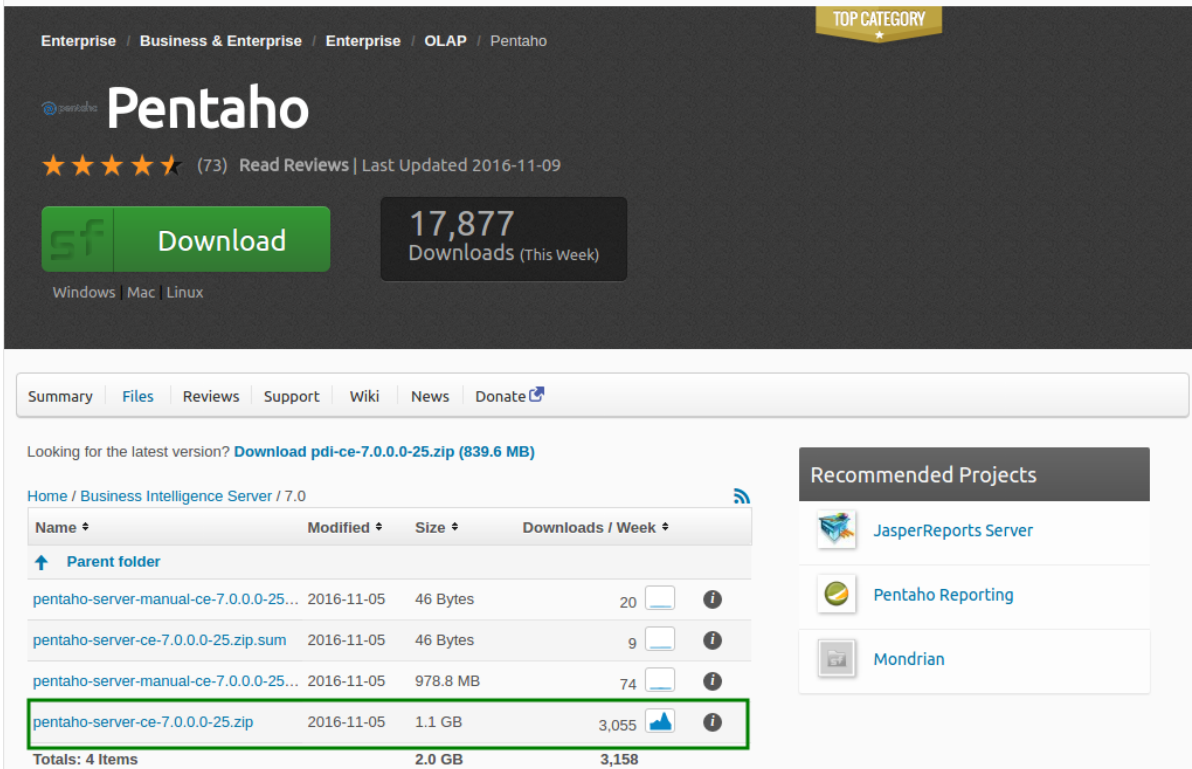
SOUSA, A. R. **Aplicação de Padrões de Projeto com a Linguagem PHP.** 2009. Monografia (Especialização em Informática) – Universidade Federal de Minas Gerais, Belo Horizonte, 2009. Disponível em:  
<<http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/BUOS-94NMN6/alissonrodrigues.pdf?sequence=1>>. Acessado em: 15 mai. 2017.

## **APÊNDICES**

## APÊNDICE A – Instalação do Pentaho Server 7.0

A ferramenta Pentaho Server fornece algumas outras ferramentas que permitem fazer análises gráficas dos dados e também navegar nos cubos. Para fazer o download da ferramenta deve-se acessar o endereço web <https://sourceforge.net/projects/pentaho/files/Business%20Intelligence%20Server/7.0/>, conforme apresentado na Figura 30.

Figura 30 - Download do Pentaho Server 7.0

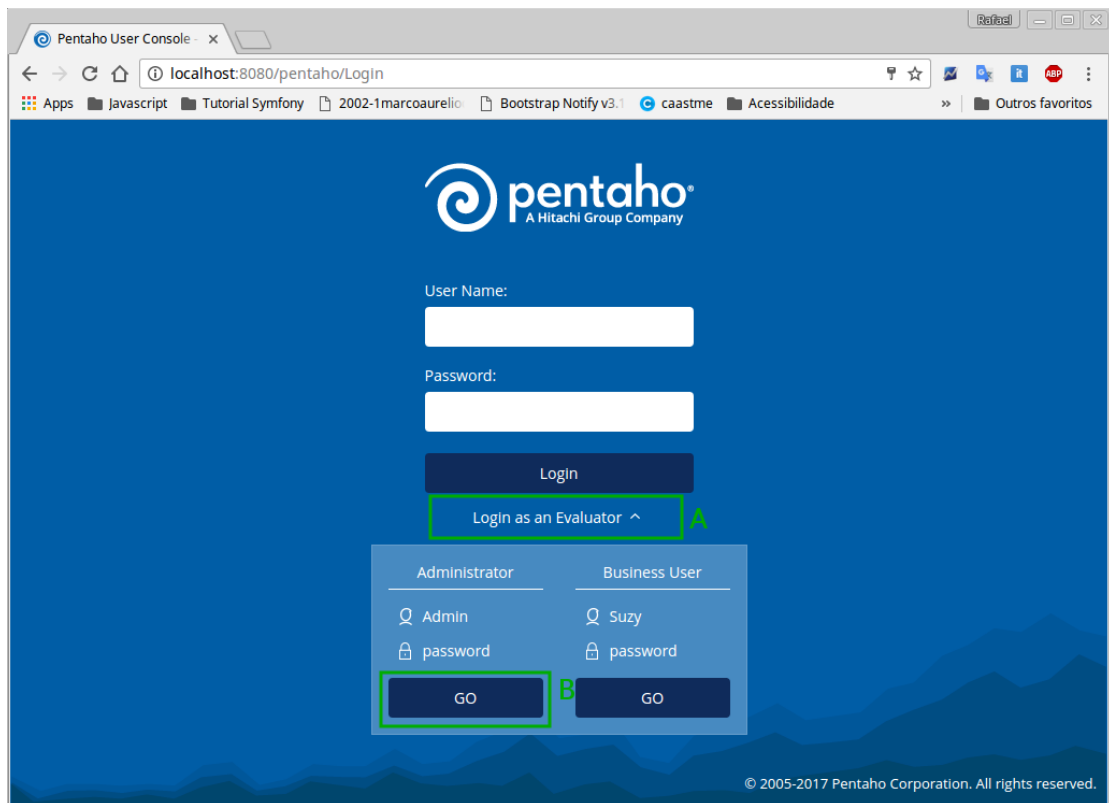


The screenshot shows the SourceForge project page for Pentaho. At the top, there are navigation links: Enterprise / Business & Enterprise / Enterprise / OLAP / Pentaho. The project name 'Pentaho' is prominently displayed with a 4.5-star rating (73 reviews) and a 'Download' button. Below the download button, it indicates '17,877 Downloads (This Week)'. A navigation bar includes links for Summary, Files, Reviews, Support, Wiki, News, and Donate. A message suggests downloading the latest version: 'Download pdi-ce-7.0.0.0-25.zip (839.6 MB)'. The main content area shows a file list table with columns for Name, Modified, Size, and Downloads / Week. The file 'pentaho-server-ce-7.0.0.0-25.zip' is highlighted with a green box. To the right, there is a 'Recommended Projects' section listing JasperReports Server, Pentaho Reporting, and Mondrian.

| Name                                    | Modified   | Size     | Downloads / Week |
|---|------------|----------|------------------|
| ↑ Parent folder                         |            |          |                  |
| pentaho-server-manual-ce-7.0.0.0-25...  | 2016-11-05 | 46 Bytes | 20               |
| pentaho-server-ce-7.0.0.0-25.zip.sum    | 2016-11-05 | 46 Bytes | 9                |
| pentaho-server-manual-ce-7.0.0.0-25...  | 2016-11-05 | 978.8 MB | 74               |
| <b>pentaho-server-ce-7.0.0.0-25.zip</b> | 2016-11-05 | 1.1 GB   | 3,055            |
| Totals: 4 Items                         |            | 2.0 GB   | 3,158            |

Na sequência selecione o link “pentaho-server-ce-7.0.0.0-25.zip” e, em seguida será exibida uma janela para fazer o download da ferramenta. Selecione o diretório que deseja salvar a ferramenta e, após o download, extraia os arquivos e execute o arquivo “start-pentaho.sh”. Uma vez que o arquivo foi executado, basta acessar o endereço web <http://localhost:8080>, que abrirá uma página de login, conforme apresentado na Figura 31.

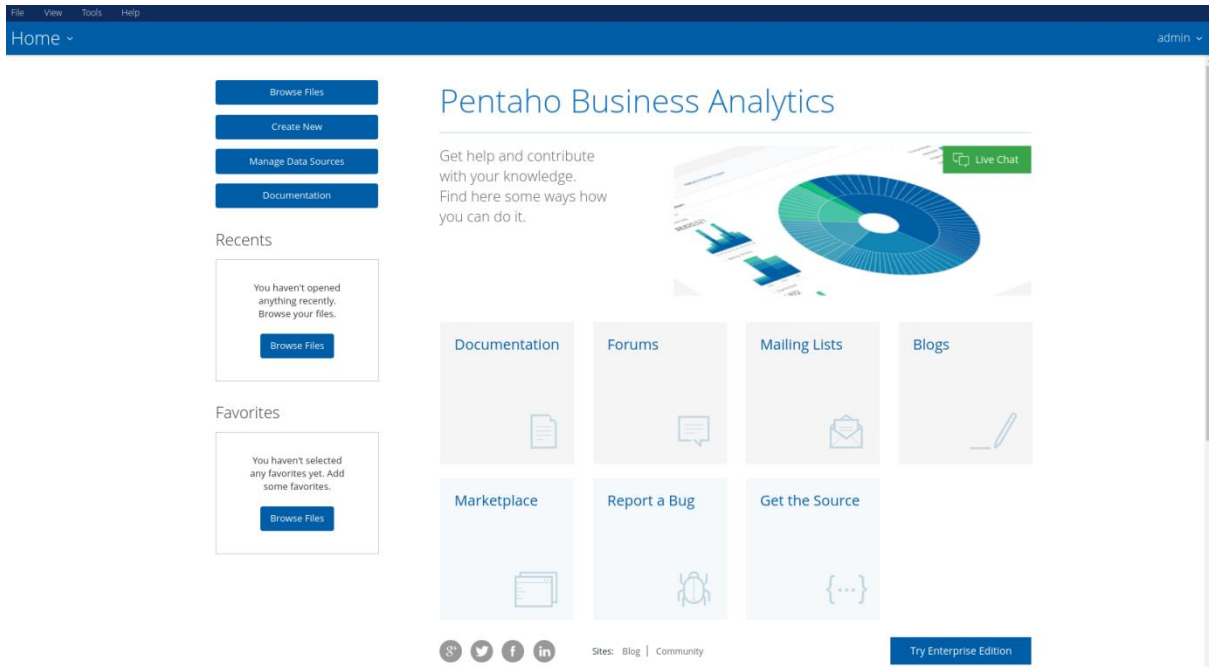
Figura 31 - Tela de login do Pentaho Server 7.0



Ao selecionar a opção “Login as an Evaluator”, conforme Figura 31-A, será disponibilizada uma conta com perfil de administrador e outra conta com perfil de usuário. Selecione a conta com perfil de administrador e a opção “GO” (Figura 31-B). A página inicial deverá ser exibida, conforme apresentado na Figura 32.



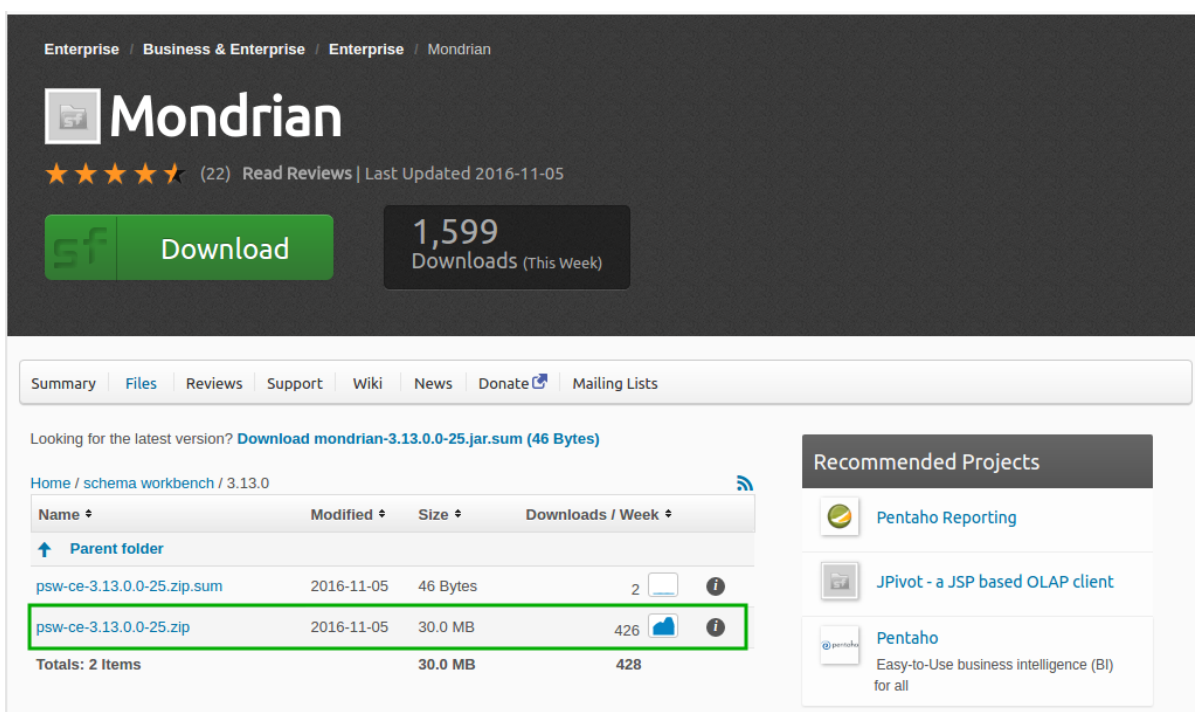
Figura 32 - Tela inicial do Pentaho Server 7.0



## APÊNDICE B – Instalação do Mondrian Schema Workbench 3.13.0

Será utilizado a versão 3.13.0 do Mondrian Schema Workbench para a criação dos cubos OLAP's. Para fazer o download da ferramenta acesse o endereço web <https://sourceforge.net/projects/mondrian/files/schema%20workbench/3.13.0>, conforme apresentado na Figura 33.

Figura 33 - Download do Mondrian Schema Workbench 3.13.0

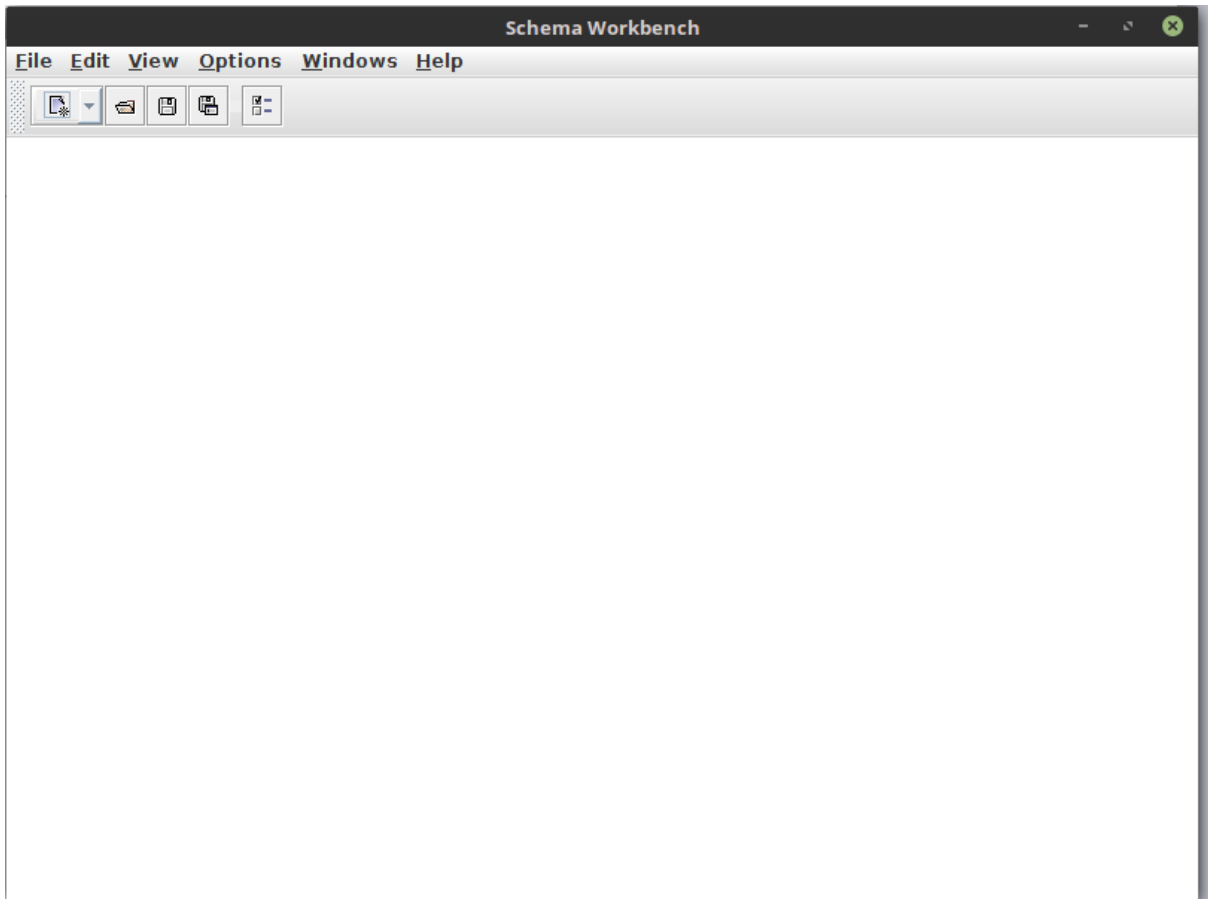


The screenshot shows the SourceForge project page for Mondrian. At the top, there is a navigation bar with 'Enterprise / Business & Enterprise / Enterprise / Mondrian'. Below this is the project name 'Mondrian' with a star rating of 4.5 (22 reviews) and a 'Last Updated 2016-11-05' date. A green 'Download' button is visible, along with a badge indicating '1,599 Downloads (This Week)'. Below the main header, there are tabs for 'Summary', 'Files', 'Reviews', 'Support', 'Wiki', 'News', 'Donate', and 'Mailing Lists'. A message says 'Looking for the latest version? Download mondrian-3.13.0.0-25.jar.sum (46 Bytes)'. The main content area shows a file list table with columns for 'Name', 'Modified', 'Size', and 'Downloads / Week'. The file 'psw-ce-3.13.0.0-25.zip' is highlighted with a green border. To the right, there is a 'Recommended Projects' section with links to 'Pentaho Reporting', 'JPivot - a JSP based OLAP client', and 'Pentaho'.

| Name                       | Modified   | Size     | Downloads / Week |
|----------------------------|------------|----------|------------------|
| ↑ Parent folder            |            |          |                  |
| psw-ce-3.13.0.0-25.zip.sum | 2016-11-05 | 46 Bytes | 2                |
| psw-ce-3.13.0.0-25.zip     | 2016-11-05 | 30.0 MB  | 426              |
| Totals: 2 Items            |            | 30.0 MB  | 428              |

Ao selecionar o link “psw-ce-3.13.0.0-25.zip” (Figura 33), é apresentada uma janela para fazer o download da ferramenta. Selecione o diretório que deseja salvar a ferramenta e, após o download, navegue até o diretório escolhido no momento do download, extraia os arquivos e execute o arquivo “workbench.sh”. Ao executar o arquivo, será exibida uma janela semelhante à apresentada na Figura 34.

Figura 34 - Tela do Mondrian Schema Workbench 3.13.0



Após a instalação da ferramenta é necessário fazer o download do drive mysql, para que seja possível configurar a conexão com o banco de dados. Para isso, acesse o endereço web <https://dev.mysql.com/downloads/connector/j/5.1.html>, conforme apresentado na Figura 35.

Figura 35 - Tela de download do drive mysql (parte 1)

Please report any bugs or inconsistencies you observe to our [Bugs Database](#).  
Thank you for your support!

**Generally Available (GA) Releases** | **Development Releases**

### Connector/J 5.1.41

Select Platform:  [Looking for previous GA versions?](#)

|   |        |      |                          |
|---|--------|------|--------------------------|
| <b>Platform Independent (Architecture Independent), Compressed TAR Archive</b><br><small>(mysql-connector-java-5.1.41.tar.gz)</small> | 5.1.41 | 3.7M | <a href="#">Download</a> |
| <b>Platform Independent (Architecture Independent), ZIP Archive</b><br><small>(mysql-connector-java-5.1.41.zip)</small>               | 5.1.41 | 4.1M | <a href="#">Download</a> |

[MD5: de520f86476298364576b541ae0f8aef](#) | [Signature](#)

[MD5: aab750c83a3eca10f8c406aae902a20](#) | [Signature](#)

[i](#) We suggest that you use the MD5 checksums and GnuPG signatures to verify the integrity of the packages you download.

Selecione a opção download como destacado na Figura 35, em seguida será apresentada outra página, como apresentado na Figura 36.

Figura 36 - Tela de download do drive mysql (parte 2)



**Begin Your Download - mysql-connector-java-5.1.41.tar.gz**

**Login Now or Sign Up for a free account.**  
An Oracle Web Account provides you with the following advantages:

- Fast access to MySQL software downloads
- Download technical White Papers and Presentations
- Post messages in the MySQL Discussion Forums
- Report and track bugs in the MySQL bug system
- Comment in the MySQL Documentation

**A** **Login »**  
using my Oracle Web account

**B** **Sign Up »**  
for an Oracle Web account

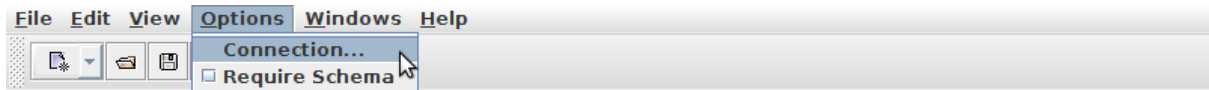
MySQL.com is using Oracle SSO for authentication. If you already have an Oracle Web account, click the Login link. Otherwise, you can sign up for a free account by clicking the Sign Up link and following the instructions.

**C** **No thanks, just start my download.**

Caso já tenha o cadastro, selecione a opção para efetuar o login, como observado na Figura 36-A. Caso seja necessário fazer o cadastro, selecione a opção apresentada na Figura 36-B. Caso seja necessário fazer apenas o download da ferramenta, selecione a opção apresentada na Figura 36-C.

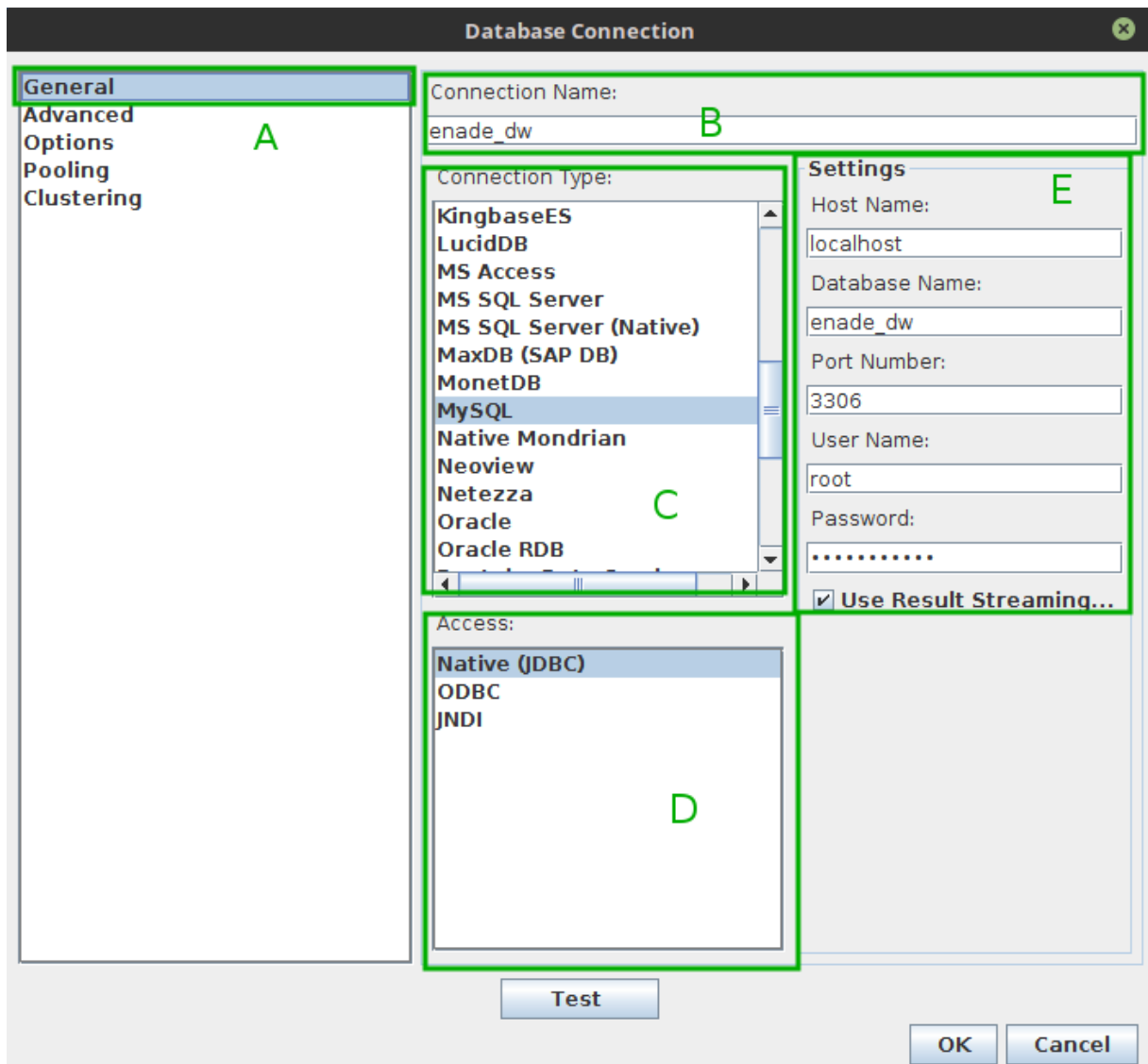
Ao selecionar a opção de download, selecione o diretório que deseja salvar a ferramenta e, assim que o download for concluído, extraia os arquivos. Após a extração, acesse a pasta extraída, copie o arquivo “mysql-connector-java-5.1.41-bin.jar” e cole-o dentro da pasta drivers que está localizada dentro do diretório /mondrian/workbench. Uma vez que o download está dentro do diretório, basta configurar a conexão no Mondrian Schema Workbench clicando na opção options/connection, como apresentado na Figura 37.

Figura 37 - Configurando a conexão do Mondrian Schema Workbench (parte 1)



Após selecionar a opção “Connection”, apresentado na Figura 37, uma nova janela será exibida, como apresentada na Figura 38.

Figura 38 - Configurando a conexão do Mondrian Schema Workbench (parte 2)

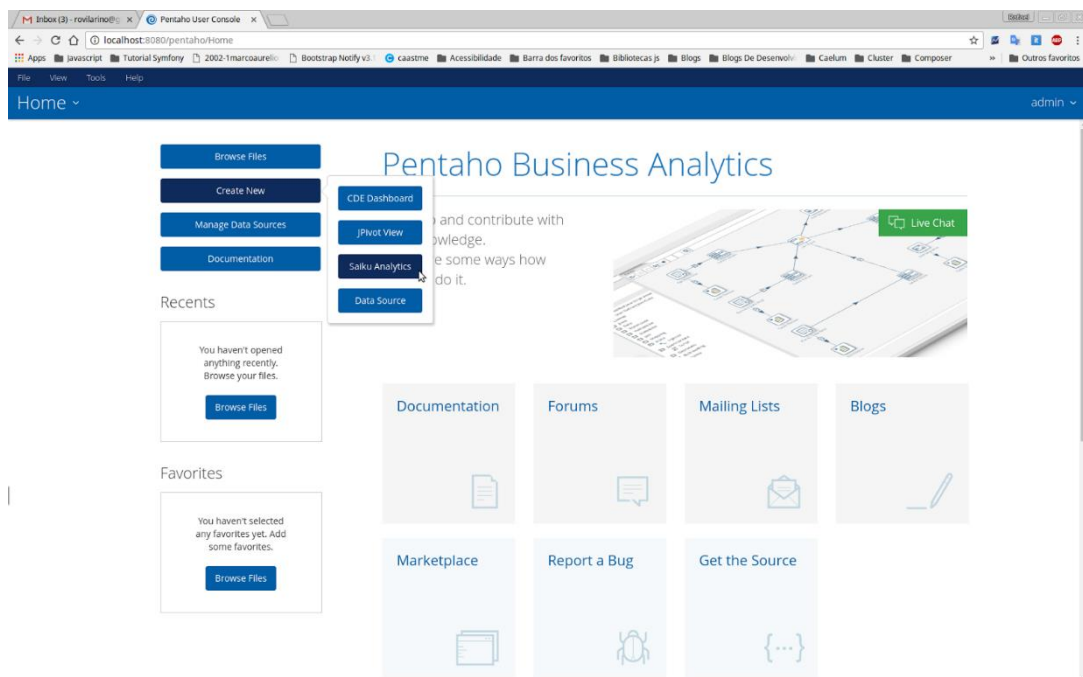


Na tela de configuração, selecione a opção “General” (Figura 38-A) e especifique um nome para a conexão na opção “Connection Name” (Figura 38-B). Na opção “Connection Type” (Figura 38-C), escolha o banco de dados utilizado, nesse caso, o MySQL, selecione o tipo de drive utilizado na seção “Access” (Figura 38-D) e informe as configurações do banco de dados na seção “Settings” (Figura 38-E).

## APÊNDICE C – Criando um gráfico no Saiku Analytics

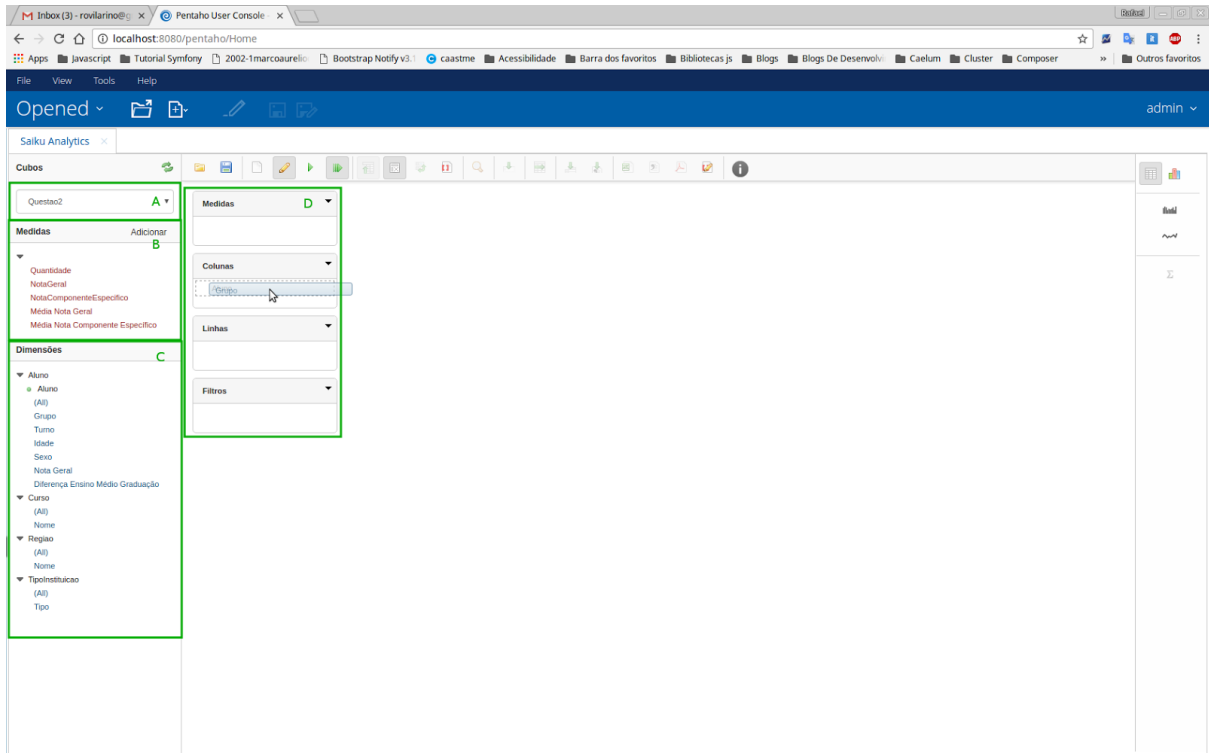
Com o Saiku é possível, criar gráficos do cubo OLAP. A medida que se deseja uma nova informação basta adicioná-la que o gráfico é atualizado. Após a publicação do cubo, basta selecionar a opção “Create New” > “Saiku Analytics”, conforme apresentado na Figura 39.

Figura 39 - Criando um gráfico no Saiku Analytics



Após abrir o plugin, o tipo “Create New Query” deverá ser selecionado. Ao selecionar a opção informada será renderizado uma nova tela na qual o cubo deverá ser selecionado, conforme apresentado na Figura 40.

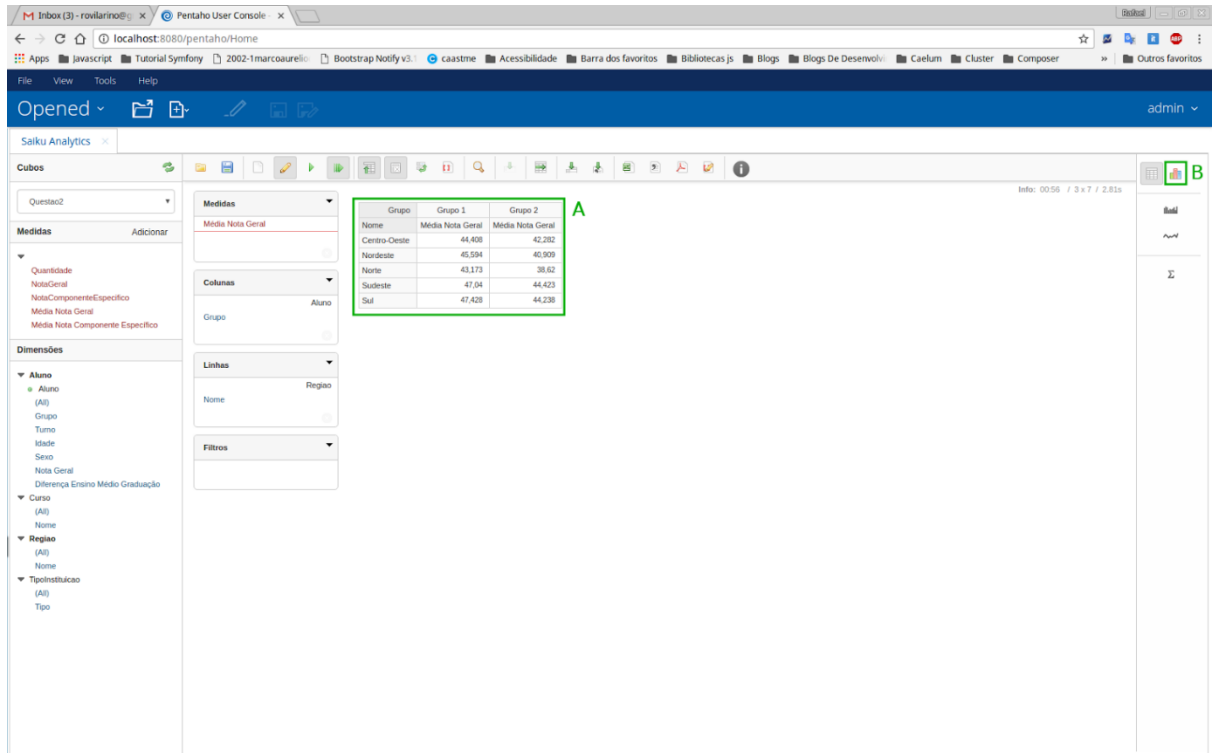
Figura 40 - Selecionando o cubo a ser trabalhado



Na Figura 40-A serão listados os cubos. Após a seleção de um cubo será exibida as “Medidas” criadas na Figura 40-B. Já na Figura 40-C serão exibidas as dimensões com suas respectivas propriedades. As propriedades das dimensões e as medidas poderão ser arrastadas para linhas, colunas, filtro e medidas de acordo com a necessidade do usuário. Ao arrastar as propriedades será exibido uma tabela com as informações adicionadas, conforme apresentado na Figura 40-D.

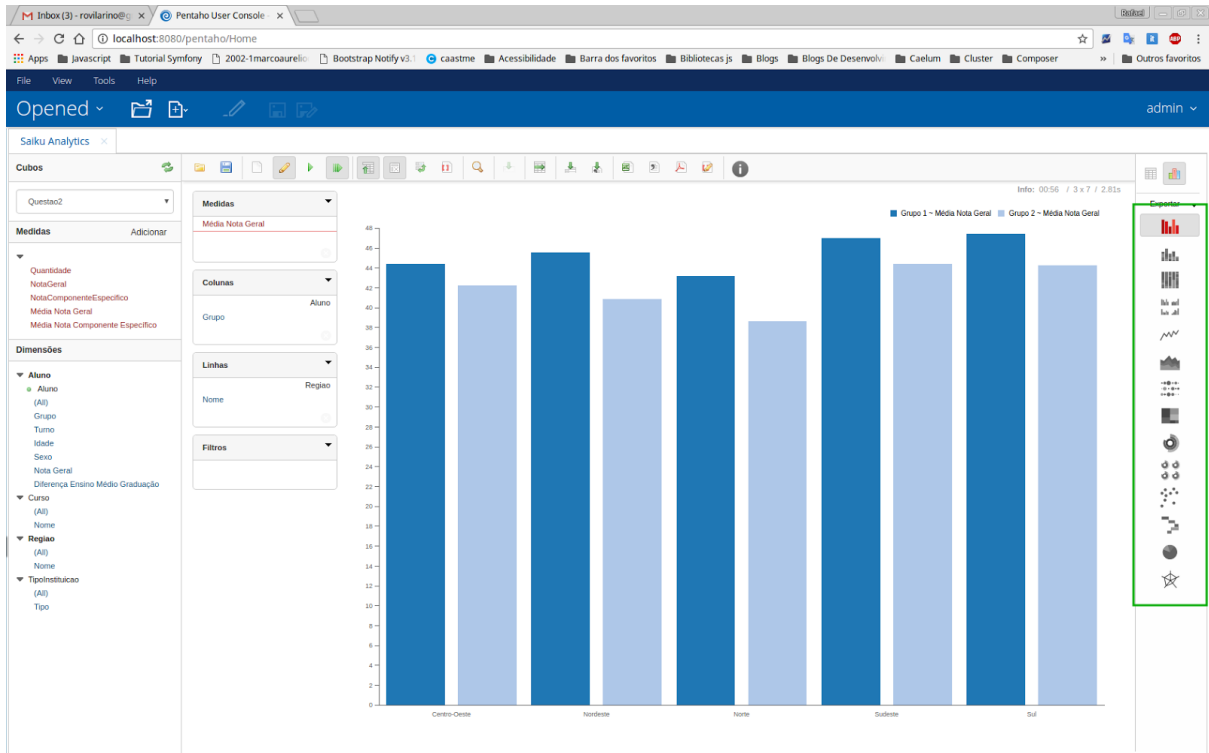


Figura 41 - Exibindo o gráfico



Para que seja exibido o gráfico, a opção representada na Figura 41-B deverá ser selecionada.

Figura 42 - Alterando o gráfico



O tipo do gráfico pode ser alterado no canto direito da ferramenta, conforme destacado na Figura 42.