

Criação de um Data Mart para análise estratégica dos concursos vestibular UFSC 2008-2012

**Giancarlo Souza de Freitas, Ronan Romeu Knob, Sabrina Schütz de Oliveira,
Valdir Luiz Hofer Arnhold**

¹Departamento de Informática e estatística(INE) – Universidade Federal de Santa Catarina (UFSC) - Florianópolis - Brasil

{giancarlosouza, ronanknob, sabrinaoliveira,
valdirarnhold}@grad.ufsc.br

Resumo. *Este artigo demonstra o processo de criação de um data Mart, para solucionar questões estratégicas de uma rede de ensino privada, utilizando dos dados das bases de candidatos que prestaram o concurso vestibular da Universidade Federal de Santa Catarina - UFSC, no período de 2008 a 2012.*

Palavras Chaves: *Data Warehouse, Data Mart, Modelo Dimensional, Banco de Dados, Vestibular UFSC, Desempenho.*

1. Introdução

Neste artigo, é demonstrada a criação de um Data Mart para solucionar questões estratégicas da administração da rede de Ensino Múltipla Escolha. Um Data Mart é uma subdivisão, ou subconjunto de um Data Warehouse, que por sua vez é um depósito de dados digitais, que armazena informações detalhadas de vários sistemas e meios, e as estrutura de forma a ajudar na tomada de decisão, tornando se assim um sistema de apoio a decisão (SAD).

O objetivo do Data Mart neste trabalho é dar suporte a uma rede privada de ensino, a rede Múltipla Escolha. Esta rede possui colégios e cursos pré-vestibular, espalhados por algumas cidades. O conselho então optou pela criação de um Data Mart para análise dos dados de alunos nos concursos vestibular da UFSC de 2008 a 2012, a fim de encontrar oportunidades de negócio, e aprimorar os serviços já oferecidos.

Este trabalho aborda os métodos para obtenção dos dados necessários, o planejamento do projeto, o projeto físico, o projeto da área de transição no Data Mart, e por fim a criação das medidas e visualização dos dados armazenados através de ferramentas de front end. Os dados utilizados advém de bases de cadastro e boletins de desempenho individual dos candidatos que prestaram o concurso vestibular da UFSC, entre os anos 2008 á 2012.

2. Materiais

Os dados usados para este trabalho provém das edições do concurso vestibular da UFSC, realizadas entre o ano de 2008 a 2012. Os dados foram fornecidos pela Comissão Permanente do Vestibular (COPERVE), órgão responsável pela organização do teste.

A inscrição para este concurso vestibular, além das perguntas usuais de cadastro, também consiste no preenchimento de um questionário sócio-econômico, onde o participante responde a perguntas, de modo a definir seu posicionamento econômico e questões sociais, o que lhe dá acesso a certos benefícios, como acesso as políticas de ações afirmativas da universidade, as quais englobam, por exemplo, o acesso a vagas reservadas a cotas reservadas a negros ou estudantes de escola pública.

As bases usadas para este trabalho ainda utilizam, além dos dados inseridos no momento do preenchimento, os resultados obtidos no vestibular. Visto isso, com esse conjunto de dados, é possível inferir e analisar correlações entre alguns aspectos sócio-econômicos, e os resultados dos participantes.

3. Métodos

Apesar de existir desde a década de 70, bancos de dados que utilizam o modelo relacional ainda são amplamente usados. A arquitetura de informações neste modelo, se baseia em abstrações de entidades do mundo real, de forma normalizada e sistemática, para facilitar o entendimento e a criação de pesquisas que conectam certas características de uma entidade com outras. Uma empresa tipicamente possui vários sistemas usados em diferentes áreas, e estes normalmente utilizam de um banco de dados relacional para armazenagem dos dados. Esses bancos de dados geralmente não estão interligados com outros sistemas.

Além do isolamento entre os bancos de dados, varrer uma grande quantidade de informações, o que é natural haver de acordo com o tempo de vida da empresa, pode ser extremamente custoso em termos computacionais. Isso é um problema quando se trata de análise estratégica, onde pode haver a necessidade de analisar grandes séries temporais.

Um tipo de sistema de apoio a decisão (SAD) que pode ajudar a resolver estes problemas, é o Data Warehouse. Um Data Warehouse (que no português significa, literalmente armazém de dados) é um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar as decisões gerenciais [1]. Neste tipo de banco de dados, os dados são organizados de forma a otimizar ao máximo o tempo de consulta, e transformar relações tipicamente encontradas em bancos relacionais, em informações facilmente entendíveis pelos usuários, o que torna fácil a utilização pelos tomadores de decisão. Utilizaremos para este trabalho um Data Mart, que é uma subdivisão de um Data Warehouse.

3.1. Definição de Data Mart

Na literatura, temos duas visões distintas a respeito da definição do Data Mart. Kimball acredita que seria um conjunto de dados relacionados a um assunto do negócio ou a um departamento da organização [2]; Enquanto Inmon acredita que o DM é uma extração de dados do DW, voltada ao atendimento das necessidades específicas de um departamento da organização [3]. Para este trabalho, utilizaremos a visão de Kimball para construção do Data Mart.

3.2. Etapas da construção do Data Warehouse/Data Mart

O projeto de um DW ou DM não difere no número de passos. Basicamente, todo o ciclo do projeto é definido em 11 passos:

- Planejamento do projeto;
- Administração do projeto;
- Definição dos requisitos de negócio;
- Modelagem dimensional;
- Projeto físico;
- Desenvolvimento e projeto da área de transição;
- Especificação da aplicação do usuário final (Front Room);
- Desenvolvimento da aplicação do usuário final (Front Room);
- Projeto e arquitetura técnica;
- Instalação e seleção de produtos;
- Implantação e manutenção.

Neste trabalho focamos no uso de 5 dos 11 passos para o alcance dos objetivos propostos deste trabalho. Detalharemos os 5 passos a seguir.

3.2.1 Planejamento do projeto

Na etapa do planejamento do projeto deste Data Mart, o escopo, justificativa, exclusões, fatores críticos de sucesso, os riscos e o plano para a abordagem destes itens foram trabalhados. Esta etapa é de importância crucial, visto que um erro no planejamento pode tornar o produto inútil ou inviabilizar o projeto.

3.2.2 Definição dos requisitos de negócio

Outra parte fundamental no plano do Data Mart, é definir quais questões estratégicas serão respondidas pelo Data Mart. Todas as etapas são importantes tecnicamente, mas essa etapa define as informações as quais todo projeto visa fornecer. Caso o projeto esteja funcional, mas não tenha os dados para responder as questões estratégicas, ele se torna inútil.

3.2.3 Modelagem Dimensional

A modelagem dimensionaonal é uma técnica de projeto lógico que procura apresentar os dados em uma estrutura padrão e intuitiva que permite o acesso de alto desempenho. É inerentemente dimensional e adere a uma disciplina que usa o modelo dimensional, mas com algumas restrições importantes. O modelo dimensional é composto de uma tabela com chave de várias partes, chamada tabela de fatos, e um conjunto de tabelas menores, denominadas tabelas de dimensão. Cada uma das tabelas de dimensão possui uma chave primária única que corresponde exatamente a um dos componentes da chave composta da tabela de fatos [4].

Existem quatro etapas para construção de um modelo dimensional, que são:

1. Definir os processos do negócio a serem modelados;
2. Definir o grão do processo do negócio. O grão é o nível fundamental atômico de dados que representará o processo na tabela de fatos;
3. Definir as dimensões que serão aplicadas a cada registro da tabela de fatos;
4. Definir os fatos mensuráveis que irão popular cada registro da tabela de fatos.

3.2.4 Projeto físico

Na etapa do projeto físico são discutidas características técnicas a respeito da infraestrutura. É ideal que neste passo seja definida a ferramenta para modelagem dos dados, e o SGBD que será utilizado no projeto. Também são definidos os tipos de dados para colunas, opções como o aceite de valores nulos. Para este aspecto, idealmente valores nulos não fazem sentido em análise dos dados. Os valores devem ser preenchidos com algum valor humano, que faça sentido numa representação gráfica, como “NENHUM” ou “SEM OPÇÃO DEFINIDA”. Também é esperada a definição das chaves primárias e secundárias e a substituição das chaves operacionais por chaves artificiais.

3.2.5 Desenvolvimento e projeto da área de transição

Nesta etapa, são definidas as regras para integração de informações das diversas fontes. Aqui são estipuladas as regras para conversões de tipos de dados e é feita a seleção de quais informações das fontes fazem parte do plano para resolver as questões estratégicas. Esse processo também é conhecido como ETL, onde cada letra representa a tarefa que é feita: Extração, Transformação e Carregamento(Loading) dos dados. As etapas são definidas a seguir:

1. **Extração** – Na extração, definem-se as origens dos dados que serão consumidos. Exemplos de origens comuns de dados podem ser: bancos de dados de sistemas de operação, planilhas, etc.
2. **Transformação** – As origens selecionadas na parte de extração, frequentemente utilizam formatos de dados próprios, pois são pensadas para o sistema ou usuário que as utilizam. O trabalho de transformação faz a uniformização desses dados, pois é crucial que os dados estejam em um mesmo formato para que possam ser agrupados corretamente. Nessa fase também é feito um importante trabalho de limpeza de dados, onde são removidas inconsistências, dados inválidos e redundâncias, o que poderia interferir nos resultados da análise.
3. **Carga** – Nesta parte, são definidos os caminhos que os dados farão, vindo das fontes selecionadas para as tabelas do DM. Geralmente são criadas tarefas automatizadas, que convertem e carregam os dados em períodos estipulados.

A ferramenta que utilizamos para a parte do ETL no nosso trabalho é o Pentaho Data Integration (ou Kettle), em sua versão 8.0. É uma ferramenta disponibilizada pela Hitachi, a empresa mantenedora, para criação de trabalhos de ETL. A ferramenta Kettle tem código aberto e pode ser utilizada gratuitamente.

3.2.6 Especificação da aplicação do usuário final (Front Room)

A parte final do nosso artigo é a apresentação gráfica dos resultados obtidos, de forma a responder as perguntas estratégicas, manipulando o cubo de dados criado. Utilizamos para isso, da ferramenta Tableau Desktop, em sua versão 10.4.1. A escolha dessa ferramenta se dá pela consolidação do nome no mercado, e da ampla documentação de apoio disponível para acelerar a obtenção dos resultados esperados.

4. Metodologia

Definidos os conceitos necessários, nesta seção detalhamos mais tecnicamente e com foco no trabalho realizado, os aspectos abordados para construção do Data Mart da rede Múltipla Escolha.

4.1 Planejamento do escopo

4.1.1 Escopo

Como mencionado na introdução, a construção do Data Mart visa descobrir, através da análise de desempenho e características socioeconômicas dos candidatos que prestaram o concurso vestibular UFSC entre 2008 e 2012, oportunidades de crescimento da rede de ensino Multipla Escolha, assim como melhorar os serviços já oferecidos pela rede.

4.1.2 Justificativa

A rede de ensino Múltipla Escolha possui em seu portfolio escolas e cursos pré-vestibular, em cidades no estado de Santa Catarina. A análise dos dados do vestibular da UFSC compreende dados fornecidos por alunos de todo o estado de Santa Catarina (além dos demais de outros estados), o que permite avaliar as necessidades e oportunidades, focadas no público alvo da rede.

4.1.3 Exclusões

O escopo desse trabalho se restringe a análise dos concursos vestibular UFSC realizados entre 2008 e 2012. Além disso, não tratamos de questões não relacionadas as questões estratégicas levantadas.

4.1.4 Riscos

Fazem parte dos riscos: Perda de dados físicos; Manipulação ou transformação errada das fontes, que pode gerar análises com resultados que não representam a realidade; trabalhar com dados não verídicos, visto que o cadastro sócio economico não é validado até a etapa de matrícula dos alunos aprovados no concurso vestibular.

4.1.4 Fatores críticos de sucesso

São fatores críticos de sucesso deste projeto: Conseguir executar as etapas necessárias de forma adequada; Manter o cronograma e entregas alinhados com o plano de projeto; Construir um esquema estrela que englobe todas as informações necessárias para resposta das questões estratégicas; Manter a equipe e as partes interessadas no projeto informadas do status do projeto e alinhar suas expectativas com os entregáveis do projeto;

4.2 Definição dos requisitos

O Data Mart elaborado neste projeto, prevê a solução das seguintes perguntas estratégicas:

- Se a posse de computadores influencia positivamente na aprovação dos alunos? (Para investir em instrumentação e recursos tecnológicos para a rede);
- Quais cidades em Santa Catarina, têm média salarial acima de 5 salários e baixa média de acertos no vestibular? (Possíveis cidades para implantar uma unidade);
- O grau de instrução da mãe e pai influencia na aprovação dos alunos? (Para avaliar a abertura de unidades para educação de adultos na rede).

4.3 Modelagem Dimensional

Com base nas perguntas, mostradas na seção 4.2, o processo que se buscou modelar é o desempenho dos alunos relacionados a suas unidades escolares e suas respostas no questionário socioeconômico, nos vestibulares da UFSC durante o período de 2008 a 2012.

4.3.1 Granularidade

Tendo como base o processo de negócio mostrado, a granularidade definida para o desempenho relacionada ao tempo é anual, isso é ela está ligada ao ano da realização do vestibular. Além disso, é possível verificar o desempenho relacionado a um curso, ou um estabelecimento de ensino, a um candidato específico, ou suas respostas do questionário socioeconômico. Relacionado ao fato, a granularidade estabelecida é atômica, registrando uma nova linha a cada interação de um aluno com o concurso vestibular.

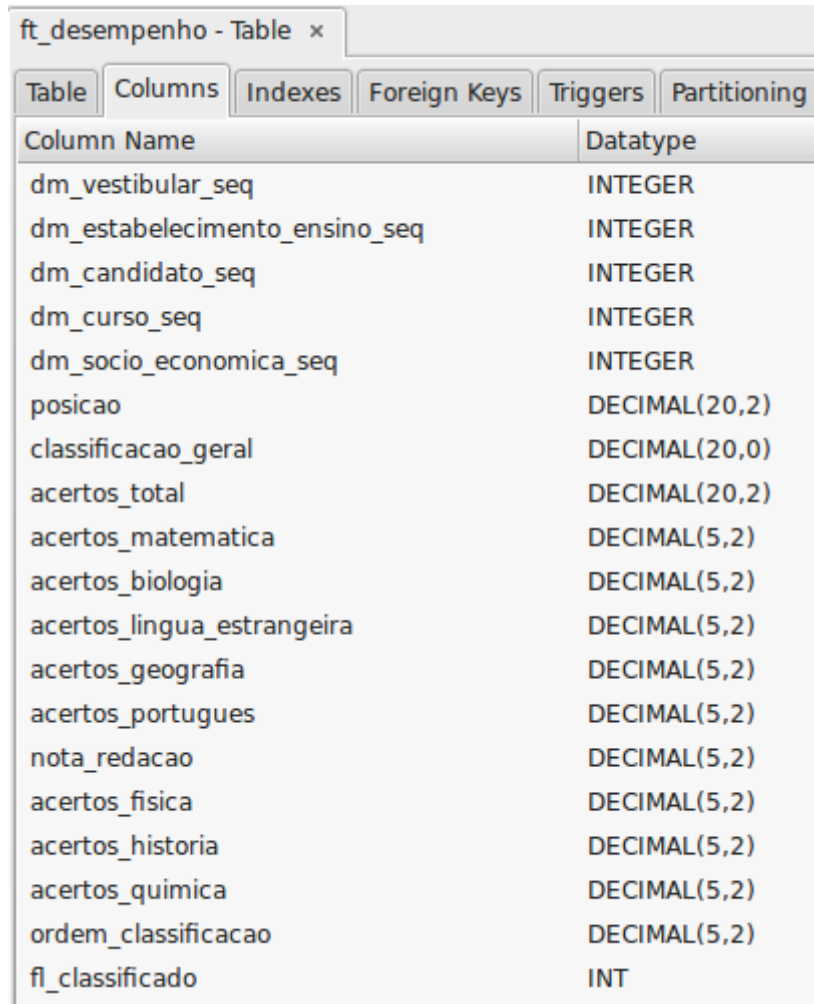
4.3.2 Dimensões

Para satisfazer os requisitos de granularidade, e o processo de negócio, foram criadas 5 dimensões que descrevem as características do desempenho que se busca medir, sendo elas:

1. Dimensão de evento: Descreve as características do vestibular, a descrição e o ano de realização.
2. Dimensão curso: Descreve as características do curso escolhido pelo candidato, o nome do curso, o nome do centro, e a área.
3. Dimensão socioeconômico: Descreve as respostas fornecidas pelo candidato, em um determinado vestibular, estão presentes na dimensão informações sobre o grau de instrução do pai, grau de instrução da mãe, se o candidato possui ou não computador, a renda total, se realizou ou não um curso pré vestibular.
4. Dimensão estabelecimento de ensino: Descreve as características do estabelecimento de ensino em que o candidato conclui o ensino médio, como o nome, a rede e a cidade.
5. Dimensão Candidato: Descreve as características de cada candidato, como a raça, a data de conclusão do segundo grau, se está realizando o concurso por experiência e se é da rede de ensino pública.

4.3.3 Fatos

O fato criado para definir a medida do processo, é o de desempenho, representado pela tabela ft_desempenho, que possui as referências das dimensões e as medidas, conforme é mostrado na figura:



Column Name	Datatype
dm_vestibular_seq	INTEGER
dm_estabelecimento_ensino_seq	INTEGER
dm_candidato_seq	INTEGER
dm_curso_seq	INTEGER
dm_socio_economica_seq	INTEGER
posicao	DECIMAL(20,2)
classificacao_geral	DECIMAL(20,0)
acertos_total	DECIMAL(20,2)
acertos_matematica	DECIMAL(5,2)
acertos_biologia	DECIMAL(5,2)
acertos_lingua_estrangeira	DECIMAL(5,2)
acertos_geografia	DECIMAL(5,2)
acertos_portugues	DECIMAL(5,2)
nota_redacao	DECIMAL(5,2)
acertos_fisica	DECIMAL(5,2)
acertos_historia	DECIMAL(5,2)
acertos_quimica	DECIMAL(5,2)
ordem_classificacao	DECIMAL(5,2)
fl_classificado	INT

Figura 1. Visão de atributos da tabela ft_desempenho

A representação completa do esquema estrela com suas dimensões ligadas ao fato é mostrada na figura:

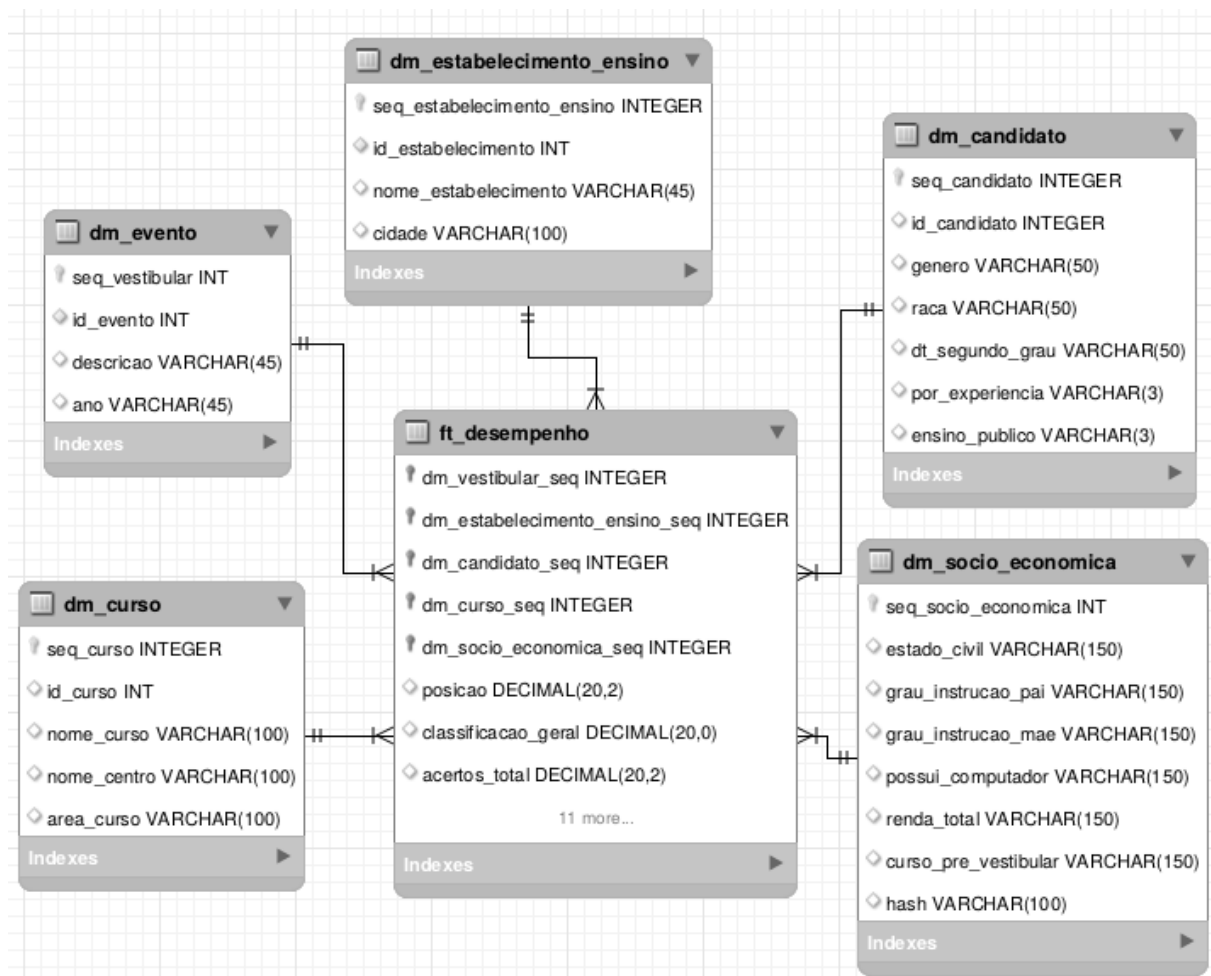


Figura 2. Visão do esquema estrela utilizado

4.4 Projeto físico

Na parte do projeto físico, foi feita a escolha do SGBD usado neste trabalho. Optou-se pelo MySQL, com a justificativa deste ser um banco amplamente conhecido, e com vasta documentação. Além disso, é uma opção grátis.

A ferramenta de suporte MySQL Workbench, na versão 6.3 CE também foi usada para manipular o cubo de dados e para criar a modelagem dimensional e fazer todas as operações de manipulação de dados necessárias.

4.4.1 Tabelas

Como pode-se observar na figura 2, seção 4.3.3, as dimensões utilizam prefixo “dm_”, enquanto a tabela de fato utiliza o prefixo “ft_”. Os títulos são compostos de letras minúsculas e refletem o conteúdo da tabela.

4.4.2 Atributos

Os atributos usados seguem o padrão de utilização, unicamente de letras minúsculas. Nomenclaturas usadas são: “id_” para ID’s verdadeiros, “seq_” para chaves artificiais nas dimensões, “dm_” para chaves artificiais que ligam as dimensões ao fato.

4.5 Desenvolvimento da Área de Transição

Para cada dimensão, foi implementada uma transformação no kettle, onde os dados são lidos, tratados e salvos no banco de dados analítico.

- candidato-transformation.ktr
- cursos-transformation.ktr
- estabelecimento-transformation.ktr
- eventos-transformation.ktr
- Socioeconomico-transformation.ktr

Todas as 5 transformações, para as tabelas de dimensões, são explicados a seguir.

4.5.1 Candidatos

A transformação dos candidatos é relativamente simples e representada na figura:



Figura 3. Transformação dos dados de candidatos

Inicialmente os dados do banco operacional são lidos no passo ‘Candidato input’, estes dados são tratados no passo ‘Ajusta Valores’ e salvos no banco analítico no passo ‘Salva Candidatos. Para o tratamento dos dados, foi utilizado um *JavaScript Parser* implementado da seguinte forma:

```

if(sexo == 'F'){
    sexoTratado = 'Feminino';
} else if(sexo == 'M'){
    sexoTratado = 'Masculino';
}

if(por_experiencia == 'N'){
    porExperienciaTratado = 'Não';
} else if(sexo == 'S'){
    porExperienciaTratado = 'Sim';
}

if(ensino_publico == '1'){
    ensinoPublicoTratado = 'Sim';
} else if(ensino_publico == '0'){
    ensinoPublicoTratado = 'Não';
}

```

Figura 4. Implementação da lógica de conversão da tabela de candidatos

4.5.2 Cursos

A transformação dos cursos foi feita em dois passos, a leitura e escrita não sendo necessário a utilização da transformação, conforme é mostrado a seguir:

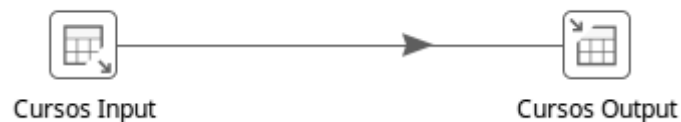


Figura 5. Conversão da tabela de cursos

A consulta SQL utilizada para a busca de candidatos foi feita da seguinte forma:

```

SELECT DISTINCT id_curso, nome_curso, nome_area, nome_centro
FROM curso
NATURAL JOIN area_curso
NATURAL JOIN centro_curso;

```

4.5.3 Estabelecimento

Para a carga da dimensão do estabelecimento também não foi necessário implementar uma transformação, conforme é mostrado:



Figura 5. Conversão da tabela de estabelecimento

A consulta SQL responsável pela leitura dos dados é a seguinte:

```
SELECT id_estabelecimento_ensino, nome_estabelecimento, rede,  
nome_municipio  
FROM estabelecimento_ensino  
NATURAL JOIN cidade;
```

4.5.4 Eventos

A carga da tabela de eventos foi efetuado em dois passos, sem conversão explícita, conforme é mostrado:



Figura 6. Conversão da tabela de eventos

A consulta para busca de eventos foi efetuada da seguinte forma:

```
SELECT descricao_evento, ano_evento, id_evento  
FROM evento
```

4.5.5 Socioeconomica

Entre todas as transformações de dimensões, a dimensão Socioeconômica demandou mais trabalho. Os passos para a transformação são mostrados na figura:

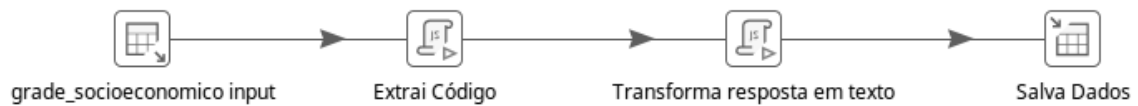


Figura 7. Conversão da tabela de dados socioeconomicos

No primeiro passo, é feita uma consulta para extrair todas as respostas dos candidatos, da seguinte forma:

```
SELECT id_evento, grade_socioeconomico
FROM candidato
```

Após isso, foi utilizado um parser, para extrair os códigos das questões definidas na tabela:

- 1.Grau instrução do pai;
- 2.Grau de instrução da mãe.
- 3.Possui computador;
- 4.Renda total;
- 5Curso de pré vestibular;

A implementação do *parser* foi feita utilizando *JavaScript*, conforme é mostrado:

```
nivel_instrucao_pai = substr(grade_socioeconomico,20,1);
nivel_instrucao_mae = substr(grade_socioeconomico,21,1);
possui_computador = substr(grade_socioeconomico,28,1);
estado_civil = substr(grade_socioeconomico,1,1);
renda_total = substr(grade_socioeconomico,19,1);
curso_pre_vestibular = substr(grade_socioeconomico,12,1);
```

Com os códigos das respostas definidos, os valores numéricos foram transformados em texto, também utilizando um parser js, e após isso os dados foram salvos na dimensão. Para realizar o *lookup* na carga do fato, foi criada uma coluna de hash na tabela.

4.6 Carga da tabela de fatos

Para implementar a carga da tabela de fatos, foi criada a seguinte transformação:

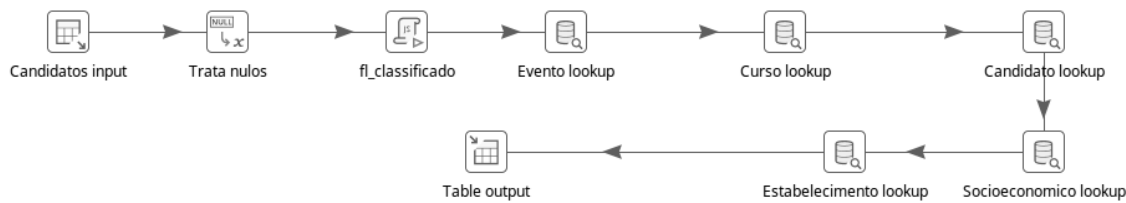


Figura 8. Esquema de carga da tabela de fatos ft_desempenho

Inicialmente é realizado um tratamento para valores nulos, da ordem de classificação e da posição, caso estes sejam nulos no banco de dados operacional, os mesmos são substituídos, por -1, para isso, foi utilizado um transformador do próprio kettle, conforme é mostrado:

Fields

Field	Replace by value	Conversion mask (Date)	Set empty str
1 ordem_classificacao	-1		N
2 posicao_1	-1		N

Figura 9. Transformador de campos utilizado na ferramenta Kettle

Após este primeiro tratamento, foi implementado um parser, para definir a flag de classificação do candidato, como 1 ou 0, conforme é mostrado:

```
var fl_classificado;
if(ordem_classificacao!= -1){
    fl_classificado = 1;
} else{
    fl_classificado = 0;
}
```

Feito isto, são feitos database *lookup* para recuperar os ids de todas as tabelas dimensões, para estes *lookups*, foram criados índices nas dimensões, o que aumentou consideravelmente a performance do processo. Como último passo, os dados são salvos na tabela de fatos.

4.7 Automação do ETL

Com o objetivo de automatizar o processo, de *Extract-Transform-Load* (ETL), foi criado um job, um recurso fornecido pelo kettle para agrupar transformações e executar rotinas, por exemplo, um enviar uma notificação ao final do processo. O Job é mostrado na figura:



Figura 10. Fluxo do Job de automação do ETL

5. Front-Room e apresentação dos resultados

Como mencionado na seção 3.2.6, a ferramenta utilizada para fazer o front-room é o Tableau Desktop. A partir dele, conseguimos gerar visualizações que nos permitiram ter uma visualização gráfica para a resposta às questões estratégicas. Abaixo, são apresentados exemplos de saída do Kettle, e seu significado quanto as questões estratégicas.

Para a questão “A posse de computadores influencia positivamente na aprovação dos alunos?”, obteve-se o seguinte gráfico em formato de pizza:

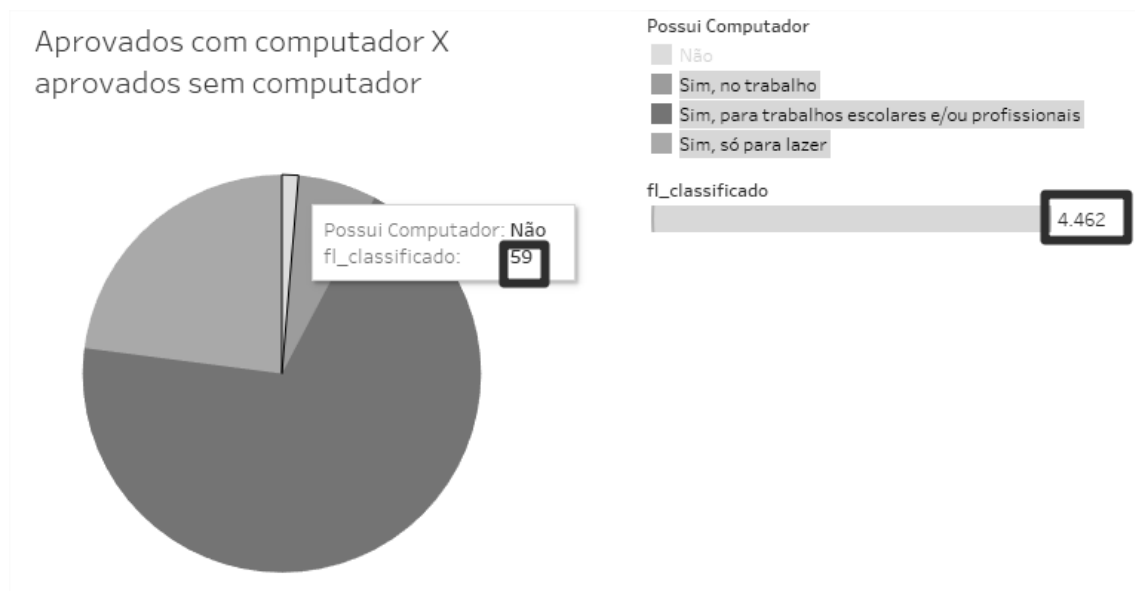


Figura 11. Visualização em formato de pizza para a pergunta estratégica 1 para o ano 2008

Na imagem, utilizamos o filtro de ano setado para mostrar apenas resultados de 2008. São destacadas duas informações: O contador de fl_classificado, que indica o número de aprovados no ano de 2008 e o número de candidatos que foram aprovados e não tem computador. Isso quer dizer que 98,68% dos candidatos aprovados possuem computador, o que indica uma relação e justifica investimento tecnológico na rede Múltipla Escolha.

Para a questão “Quais cidades em Santa Catarina, têm média salarial acima de 5 salários e baixa média de acertos no vestibular?”, trabalhamos com alguns filtros. Primeiro, levantamos a média de acertos dos candidatos aprovados. Esse dado pode ser obtido com facilidade através das páginas na internet, referentes ao vestibular de cada ano. Porém, como temos acesso aos dados, levantamos por meio da seguinte pesquisa:

```
SELECT AVG(acertos_total)
FROM ft_desempenho
WHERE fl_classificado = 1
```

Feito isso, levantamos que a média de acerto dos candidatos aprovados entre 2008 e 2012 foi 56.5 pontos. A partir disso, no Tableau, selecionamos também um filtro de renda, pois desejamos só filtrar cidades com alta média salarial:

Filtro [Renda Total]

Geral Curinga Condição Superior(es)

☒ Selecionar na lista ☐ Lista de valores personalizados

Inserir texto de pesquisa

- ☐ Acima de 1 até 3 sal. mín.
- ☐ Acima de 3 até 5 sal. mín.
- ☒ Acima de 5 até 7 sal. mín.
- ☒ Acima de 7 até 10 sal. mín.
- ☒ Acima de 30 sal. mín.
- ☐ Até 1 salário mín.
- ☒ Entre 10 e 20 sal. mín.
- ☒ Entre 20 e 30 sal. mín.

Figura 12. Filtro do atributo renda_total no Tableau

Utilizamos então para construção da visualização, cidades que possuem candidatos com alta média salarial, mas que obtiveram uma baixa média de acertos totais no vestibular (até 35 acertos). O resultado foi de 17 cidades que correspondem com estes aspectos, como apresentado na visualização de mapa, gerada para análise desta resposta, mostrado abaixo:



Figura 13. Visualização da resposta a pergunta estratégica 2

Por último, para a questão “O grau de instrução da mãe e pai influencia na aprovação dos alunos?”, utilizamos a visualização do tipo bolhas em pacotes. Como o grau de instrução da mãe e do pai são duas informações fornecidas separadamente no questionário socioeconômico, tivemos de avaliar separadamente. Abaixo os resultados:

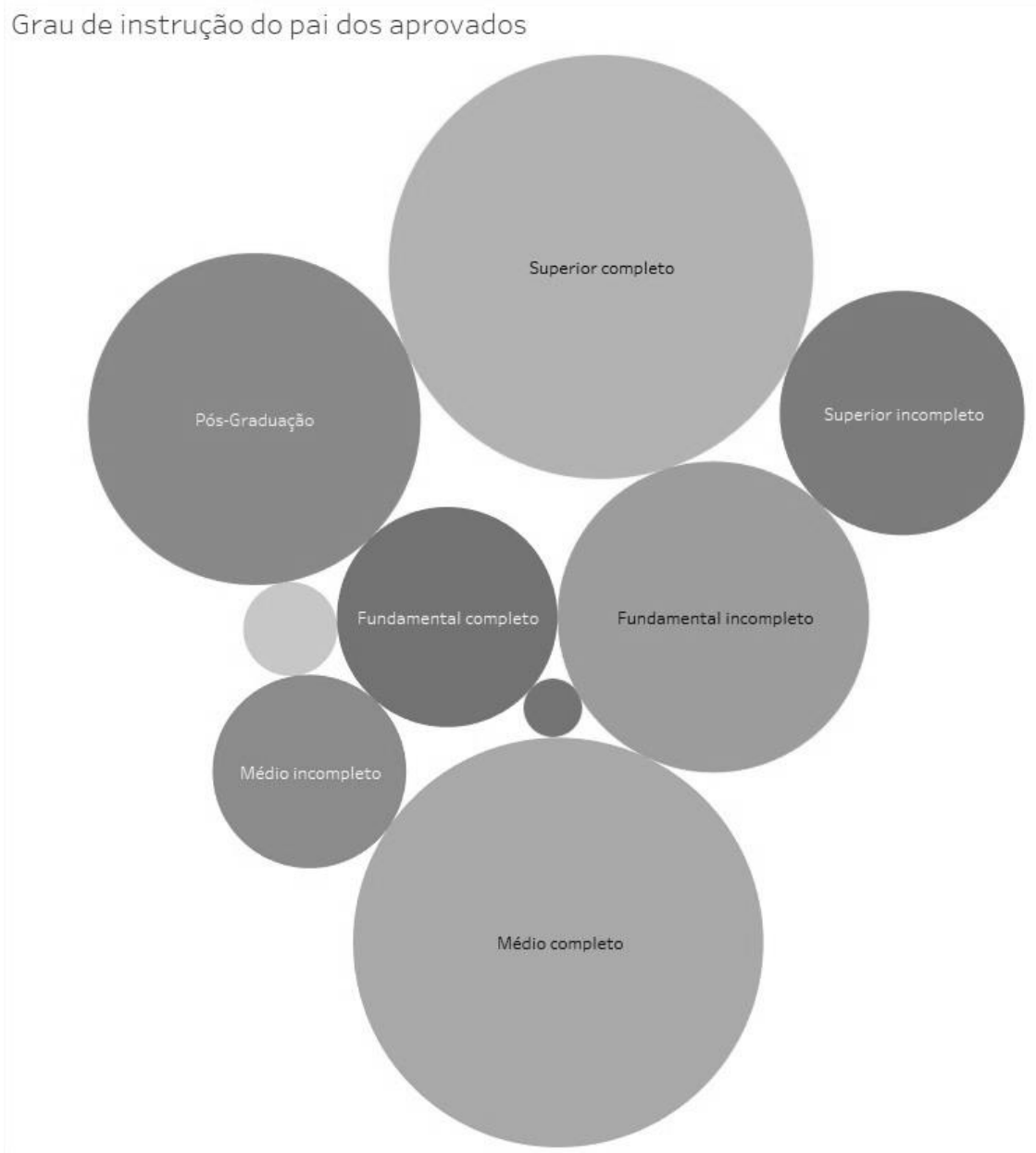


Figura 14. Visualização da relação de grau de instrução do pai dos aprovados

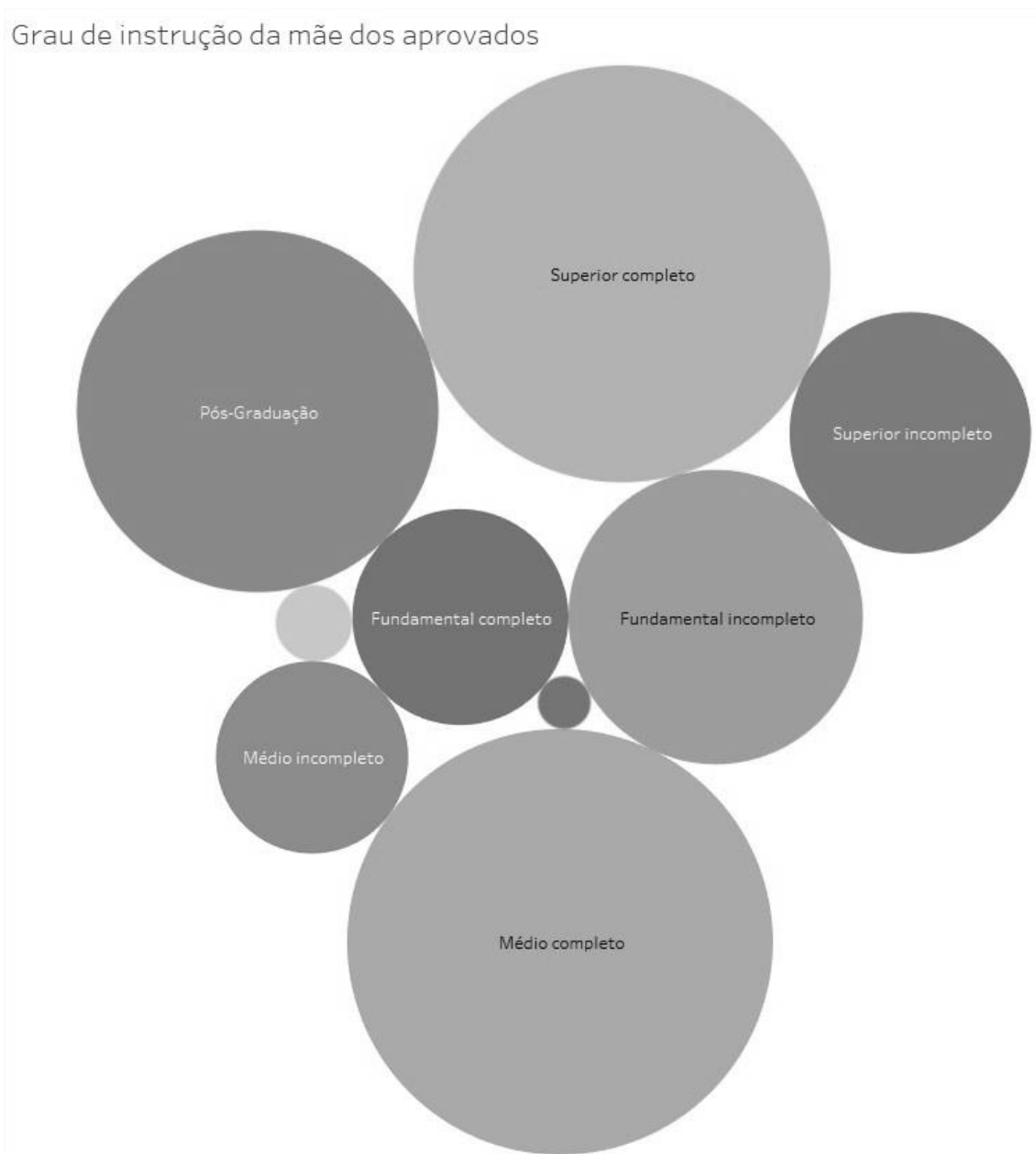


Figura 15. Visualização da relação de grau de instrução da mãe dos aprovados

Com a análise das duas visualizações, pode-se notar dois fatos: Que a maior parte dos aprovados no vestibular, tem pais com ensino médio completo ou superior, o que pode ser um indicativo da influência da educação dos pais na aprovação dos filhos. Além disso, nota-se também que na classificação da educação de pais e mães, o padrão é muito parecido, o que indica que há praticamente a mesma porcentagem de mães e pais com o mesmo nível de educação.

6. Conclusões

Neste trabalho foi possível aplicar os conceitos relacionados a criação de um Data Mart, que também podem ser aplicados na criação de um Data Warehouse. Observamos por meio desse o poder da solução em oferecer, de modo confiável, dados para auxiliar na tomada de decisão, objetivo principal de um projeto de DW.

No exemplo apresentado, foi possível estruturar um esquema estrela que proveu todos os dados necessários para solução das questões estratégicas levantadas, cumprindo seu objetivo. Além disso, a visualização no front-room nos permitiu montar facilmente gráficos para solução das questões estratégicas, o que mostra que este esquema estrela poderia ser usado pelo nível estratégico de uma empresa, o qual era o objetivo do trabalho.

Referências

- [1] DATE, C. J. Introdução a Sistemas de Bancos de Dados. 8ª Ed., Rio de Janeiro: Campus, 2004.
- [2] HACKATHORN, R. D.; INMON, W. H. Using the data warehouse. [S.l.]: Wiley-QED Publishing Somerset, 1994. ISBN 0-471-05966-8.
- [3] KIMBALL, R.;ROSS, M. The data warehouse toolkit: the complete guide to dimensional modelling. New York [ua]: Wiley, 2002.
- [4] KIMBALL, R. (1997). "A dimensional modeling manifesto". Disponível em <<http://www.kimballgroup.com/1997/08/a-dimensional-modeling-manifesto/>>. Acesso em 20 nov. 2017.