Universidade Federal de Santa Catarina

Projeto Final — Catherine

Data Warehouse — INE5643

Augusto Fredigo Hack
Fabricio Bez
Luis Felipe Nunes
Matheus Hoffmann Silva

Florianópolis, 26 de julho de 2013

Conteúdo

1 Resumo

TODO
Projeto Catherine. (O quê? Como? O que resultou? ¿ Palavras-chave)

2 Introdução

As informações coletadas anualmente pela Universidade Federal de Santa Catarina com as inscrições e resultados de seu vestibular proporcionam uma ampla base de dados. Esta base contém informações sobre todos os candidatos que prestaram vestibular, tanto dados sócio-econômicos (provenientes do questionário sócio-econômico aplicado no momento da insrição do vestibular) e dados como endereço, data de nascimento do candidato, quanto resultados das provas.

O projeto Catherine procura utilizar os dados mencionados acima para analizar a situação das cidades do Estado de Santa Catarina quanto à média de aprovações no vestibular, a fim de identificar as regiões com menor média de aprovação. Estas regiões podem ser o foco de atuação de empresas que vendem Cursos Preparatórios para o vestibular, visto a defasagem da região, que podem ser um bom nicho de negócio.

Para enriquecer a base de dados já existente, será usado um levantamento de dados estatístico-educacionais disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Desse levantamento geral, serão utilizados os dados referentes às instituições de ensino por cidade de Santa Catarina. Estes dados serão úteis para saber onde há regiões onde os estudantes não prestaram vestibular para a UFSC, além de ter dados sobre quantos estudantes se formam e quantos efetivamente prestam vestibular.

3 Materiais

Os dados de candidatos que prestaram vestibular foram disponibilizados pela Comissão Permanente do Vestibular (Coperve), que é a responsável por realizar o concurso vestibular para ingresso na UFSC. Para preservar a identidade e zelar pela privacidade dos indivíduos cujas informações estão presentes no banco de dados, não há meios de identificação presentes no banco, como nome, registro de identidade ou carteira de pessoa física.

No caso deste trabalho, serão utilizados os dados coletados dos vestibulares da UFSC para os anos de 2008 a 2012. Estes dados estão disponíveis em formato .sql, que é um *dump* do banco de dados MySQL.

Além do Banco de Dados, está disponível o modelo Entidade Relacionamento e uma documentação on-line com mais detalhes do banco, que facilitam o entendimento do mesmo.

4 Métodos

TODO
(Revisão sobre DW e Data Mart, vislumbrando esta tecnologia como a solução para o problema)

5 Metodologia

A metodologia utilizada neste projeto é a proposta por Ralph Kimball (Modelagem de Dimensões).

5.1 Escopo

O projeto Catherine visa utilizar os dados dos vestibulares UFSC de 2008 a 2012, disponiblizados pela Coperve, para analizar a situação das cidades do Estado de Santa Catarina quanto à média de aprovações no vestibular, com o objetivo de identificar as regiões com menor média de aprovação. A identificação destas regiões podem ser úteis, no caso de empresas que vendem Cursos Preparatórios para o vestibular, para investir em novas áreas.

Além dos dados da Coperve, serão utilizados dados do INEP, que indicam o número de estudantes, das várias fases de ensino, das cidades de Santa Catarina.

5.2 Justificativa

A criação do Data Warehouse se justifica pela complexidade das *queries* no caso de utilizar um modelo de banco de dados relacional (convencional), além do uso de dados externos ao banco de dados disponibilizados pela Coperve.

5.3 Exclusões

Não fazem parte do foco do projeto Catherine, ainda que disponíveis nas bases de dados da Coperve:

- Dados de candidatos que fizeram a prova como treino.
- Dados de candidatos faltantes.
- Informações de opções (1, 1a ou 2) de cursos de candidatos ou da escolha da língua estrangeira.

5.4 Riscos

TODO

5.5 Fataores Críticos de Sucesso

TODO

5.6 Definição dos Requisitos

TODO

5.7 Modelagem Dimensional

5.7.1 Definição do Processo a ser Modelado

A modelagem estrela vai envolver os dados referentes às provas e resultados, locais de origem dos candidatos, data data dos eventos e censo referente às escolas de Santa Catarina.

5.7.2 Definição da Granularidade

Conforme o rítmo anual dos vestibulares da UFSC, foi definido a granularidade de tempo em **anos**. A granularidade de local, é em **cidades**.

5.7.3 Definição das Dimensões

As dimensões identificadas no modelo estrela foram:

- Tempo: contém as indicações de tempo tanto em números quanto em textos, do ano, mês e dia.
- Local: indicação geográfica dos eventos, precisando a cidade e estado (por extenso e UF).
- Curso: O nome do curso, nome e sigla do centro onde é oferecido o curso.

5.7.4 Definição dos Fatos

Foram criados os seguintes Fatos para o Data Mart do projeto Catherine:

- Censo: contém os principais indicadores do projeto, **por cidade**: média de acertos e notas dos candidatos, percentual de aprovação dos candidatos, se a maioria dos candidatos cursavam algum pré-vestibular, a quantidade de candidatos que cursou pré-vestibular e também a renda familiar bruta média dos candidatos. Além disso, números referentes à quantidades de estudantes que concluíram o ensino médio, por ano, para cada cidade do país.
- Vagas: refere-se às informações de vagas de cada curso. Contém o curso, ano das medidas, número de vagas oferecidas, médias de acertos e notas de redação.

5.7.5 Modelo

O modelo dimensional é apresentado abaixo:

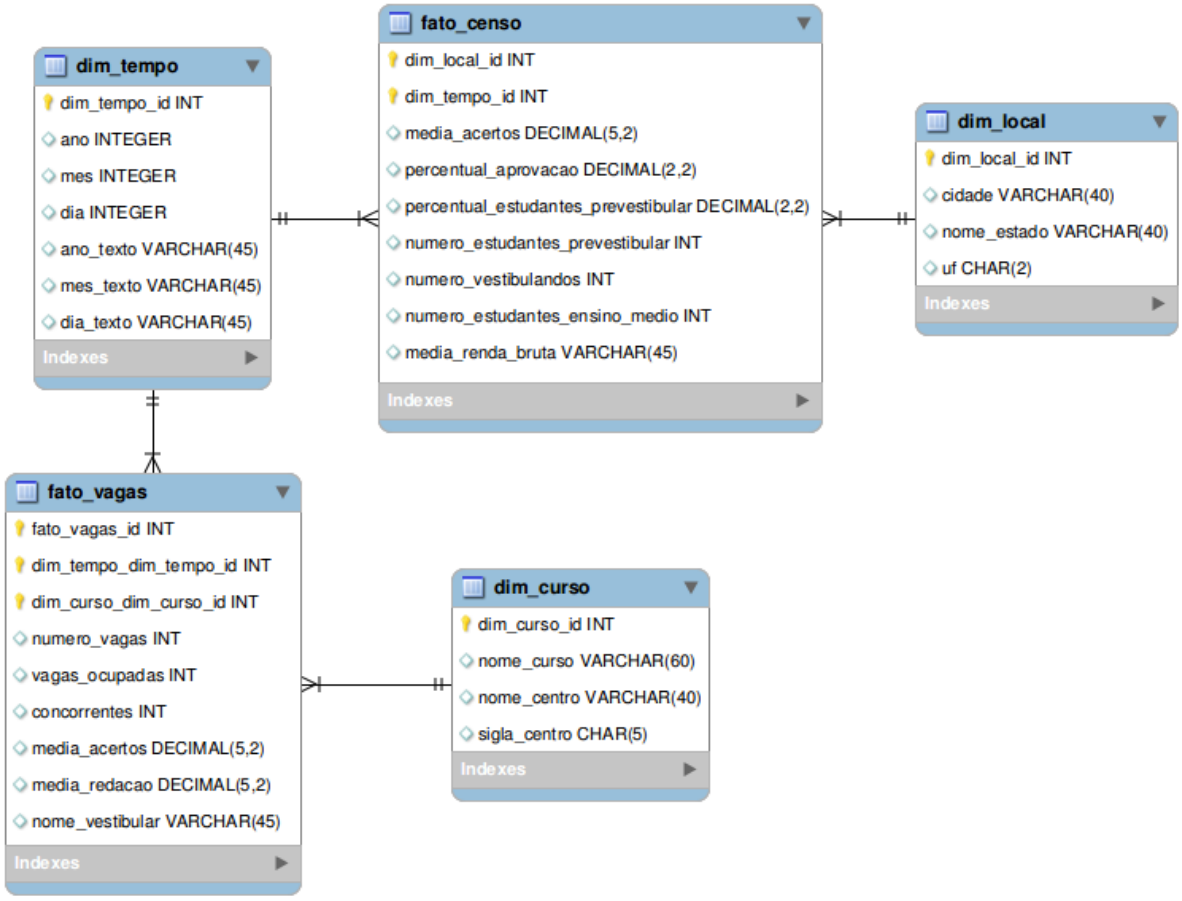


Diagrama gerado utilizando a ferramenta MYSQL WORKBENCH.

5.8 Projeto Físico

O projeto físico foi desenvolvido utilizando a ferramenta de gerenciamento de bancos de dados MYSQL, onde estão armazenados os dados do *Data Mart*. O modelo dimensional foi gerado com a ferramenta MYSQL WORKBENCH. Quanto à padronização do modelo físico, foram adotados os seguintes padrões:

- Os nomes das tabelas e campos contém apenas letras minúsculas.
- Tabelas e campos com mais de uma palavra, devem usar '_'para dividir as palavras.
- O nome de uma dimenção deve começar com 'dim_', e o nome de um fato deve começar com 'fato_'.

5.9 Projeto da Área de Transição – ETL

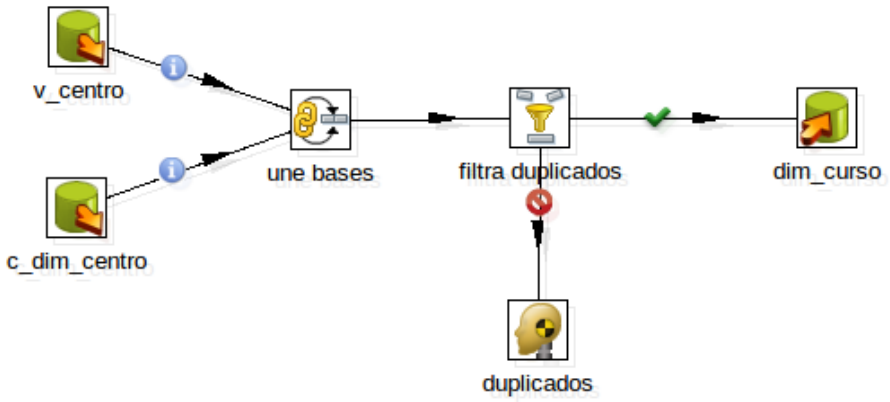
O projeto da àrea de transição foi realizado utilizando a ferramenta Pentaho Data Integration (Kettle). Esta ferramenta utiliza-se dos seguintes conceitos principais: transformations e jobs.

As transformations são um conjunto de steps que são as atividades relacionadas a integração de dados, como: carregamento de dados de uma tabela, operações com strings, filtros de valores, etc. Na transformation os *steps* são executados em paralelo. Os jobs servem para estabelecer um fluxo na execução das transformations de maneira serial.

Neste trabalho a geração de cada fato e cada dimensão foi separada em transformations e a ordem de execução das transformations em jobs.

5.9.1 Área de Transição

A primeira base de dados a ser gerada foi a da dimensão **Curso**. Esta base tem geração simples, e é apenas necessário informações das tabelas centro e curso do modelo relacional. Abaixo segue a *transformation* da dimensão Curso:

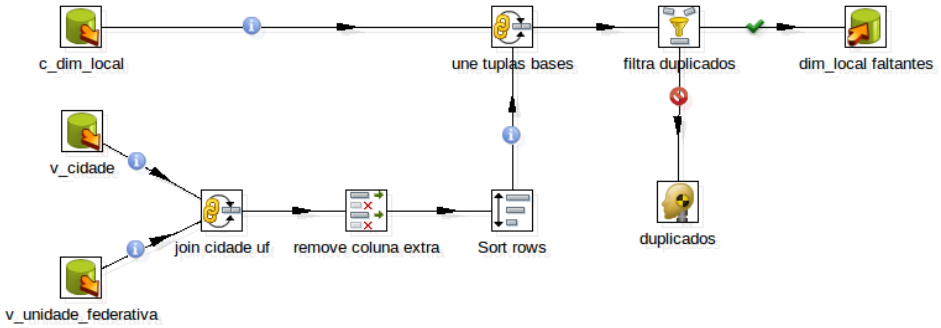


Transformation da dimensão Curso.

Os passos ilustrados acima são:

- Entrada de dados: dois fluxos, um coletando informações referente ao curso e outro referente ao centro.
- União dos dados: *join* dos dois fluxos anteriores.
- Filtragem de duplicados: filtra tuplas com informações duplicadas, gerando o *output* **dim_curso**.

A seguir, foi gerada a dimensão **Local**. Além de utilizar as tabelas Cidade e Unidade Federativa do modelo relacional, foram acrescentadas todas as outras cidades que, por ventura, não estejam contidas no banco de dados da Coperve. A *transformation* para a dimensão Local segue abaixo:



Transformation da dimensão Local.

Procedimento de criação da dimensão Local:

- Existem três fluxos de importação de dados: dois fluxos provenientes do banco de dados disponibilizado pela Coperve, e outro com dados retirados do governo, que contém nomes das cidades brasileiras.
- Os dados provenientes do banco de dados da Coperve são selecionados e sofrem uma união (join cidade uf) e, após retirar dados duplicados (remove coluna extra), são ordenados (Sort rows).
- Então as duas fontes de dados são unidas (une tuplas bases), e os dados duplicados são removidos (filtra duplicados).
- Ao final do processo, gera-se o *output* **dim_local**.

A dimensão **Tempo** foi gerada utilizando *bash script*, por questão de eficiência. Esta dimensão contém datas de 01 de janeiro de 2000, até 01 de janeiro de 2021. Os dados estão divididos entre dia, mês e ano (valores numéricos) e mês e dia da semana (texto).

- Colocar passos/bash aqui.

Para a criação do fato **Censo**, além de utilizar os dados do modelo relacional disponibilizado pela Coperve, faz uso de dados provenientes do censo escolar disponibilizado pelo INEP, que indicam a quantidade de estudantes, por cidade, que finalizaram o ensino médio (escolas públicas e privadas). Para tal, como ilustrado na *transformation*, existem dois fluxos principais de entrata de dados.

Transformation do fato Censo.

O fato **Vagas** é gerado utilizando apenas os dados disponibilizados pela Coperve. Segue a *transformation*:

Transformation do fato Vagas.

TODO: Jobs: 1-Deploy Dimensões; 2-Deploy Fatos; 3-Deploy geral.

6 Resultados

TODO

7 Conclusão e Trabalhos Futuros

TODO

8 Bibliografia

TODO

Florianópolis, 26 de julho de 2013.