

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	2
2	Introdução	2
3	Materiais	2
4	Métodos	2
5	Metodologia	2
5.1	Escopo	3
5.2	Justificativa	3
5.3	Exclusões	3
5.4	Riscos	3
5.5	Fatores Críticos de Sucesso	3
5.6	Definição dos Requisitos	3
5.7	Modelagem Dimensional	3
5.7.1	Definição do Processo a ser Modelado	3
5.7.2	Definição da Granularidade	3
5.7.3	Definição das Dimensões	3
5.7.4	Definição dos Fatos	4
5.7.5	Modelo	5
5.8	Projeto Físico	5
5.9	Projeto da Área de Transição – ETL	6
5.9.1	Preparação da Base de Dados	6
5.9.2	Área de Transição	6
6	Resultados	6
7	Conclusão e Trabalhos Futuros	6
8	Bibliografia	6

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 inscrição 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 analisar 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, disponibilizados pela Coperve, para analisar 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 Fatores 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 dos eventos e censo referente às escolas de Santa Catarina.

5.7.2 Definição da Granularidade

Conforme o ritmo 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, e também a classe social predominante dos candidatos. Além disso, números referentes à quantidades de estudantes que concluíram o ensino médio, por ano, para cada cidade do país.
- Prova: este fato representa uma prova (na verdade o conjunto das provas do vestibular) realizada por um único candidato. Contém os acertos de todas as matérias e nota da redação, classificação geral e por opção, curso prestado, local de origem do candidato, ano da prova, e ainda as respostas do questionário sócio-econômico, que tem relação com cursos preparatórios.
- 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:

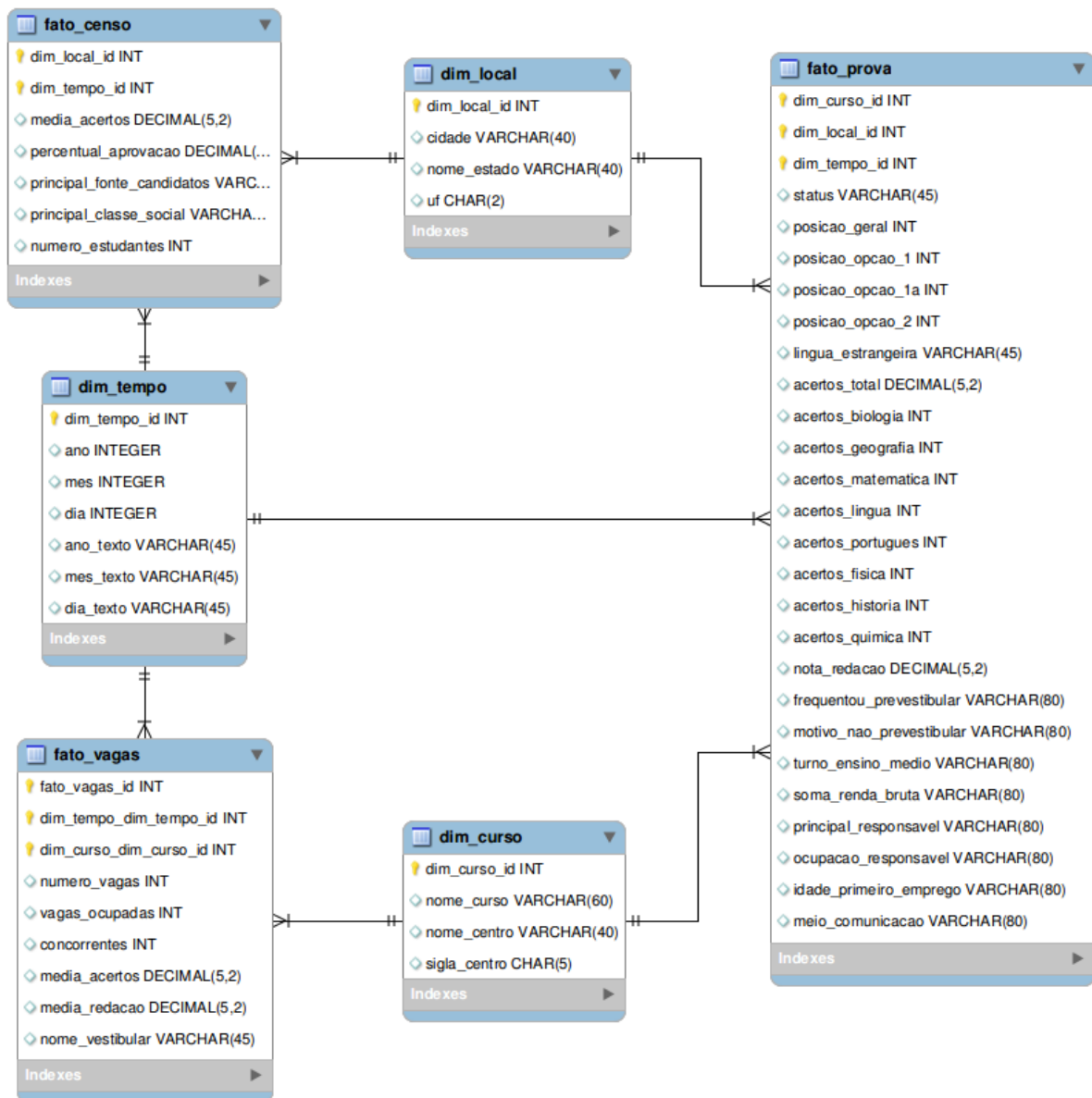


Diagrama gerado utilizando a ferramenta MYSQL WORKBENCH.

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 fer-

ramenta 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 dimensão 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 área de transição foi realizado utilizando a ferramenta Kettle, da Pentaho. 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 Preparação da Base de Dados

TODO

5.9.2 Área de Transição

TODO

6 Resultados

TODO

7 Conclusão e Trabalhos Futuros

TODO

8 Bibliografia

TODO