Universidade Federal de Santa Catarina -Campus Florianópolis
Departamento de Informática e Estatística – Universidade Federal de Santa Catarina
Disciplina de Data Warehouse - Curso de Sistemas de Informação
Prof. Dr. José Leomar Todesco

Data Mart para análise dos dados dos vestibulares da Universidade Federal de Santa Catarina entre os anos de 2008 a 2012

Diogo Silva Bach, Otávio Augusto Corrêa, Thiago Diniz da Silveira
Florianópolis, SC – Brasil
Julho de 2017

Resumo

O artigo aborda etapas de planejamento e desenvolvimento de um projeto de Data Mart para estudo de dados dos candidatos do vestibular da Universidade Federal de Santa Catarina. As análises previstas neste trabalho têm o objetivo de fornecer informações sobre a relação entre os candidatos, sua região e seu desempenho e o curso escolhido, assim como informações socioeconômicas a respeito dos candidatos.

1. INTRODUÇÃO

Neste artigo será apresentado um estudo de caso para realização de um Data Mart para análise dos dados dos vestibulares UFSC de 2008 a 2012, com finalidade de responder algumas perguntas gerenciais sobre o perfil dos candidatos e sua situação socioeconômicas. Com isto é possível analisar quais foram as dificuldades, os problemas e as diferenças entre as classes, regiões, etc.

Faz parte do trabalho apresentação dos principais conceitos e etapas para a construção do Data Mart utilizado para a produção do projeto, além do planejamento, projeto, modelagem dimensional e perguntas estratégicas sobre os dados do esquema de dados.

2. DADOS UTILIZADOS

Foi utilizado neste trabalho, uma base de dados dos candidatos que prestaram vestibular da Universidade Federal de Santa Catarina entre os anos de 2008 e 2012.

A base de dados contém informações sobre a situação socioeconômica dos candidatos, bem como opções de cursos e aprovações.

3. MÉTODOS

Para execução dos estudos necessários para este trabalho foi desenvolvido um esquema de Data Mart, considerando os dados obtidos e a análise feita sobre os mesmo. Data Mart é um subconjunto de um Data Warehouse.

Data Warehouse tem por definição literal um depósito de dados onde é possível a guarda de informações e dados da atividade da organização em banco de dados, possui ainda características de não volatilidade, variável em relação ao tempo, e é constituído de forma integrada.

As etapas realizadas neste trabalho foram:

1. Planejamento do projeto: Nessa etapa foi definido o escopo, as justificativas, riscos, alguns dos fatores críticos de sucesso e tudo aquilo que considera-se exclusões. Também podemos dizer que essa etapa ocorre a iniciação do projeto.

2. Definição dos Requisitos de Negócio: São identificadas as necessidades informacionais da organização e deve-se definir as perguntas estratégias que o Data Mart deve responder.

3. Modelagem Dimensional: Através de técnicas de desenho e de apresentação de dados de forma padronizada, são produzidas as dimensões.

4. Projeto Físico : Aqui é selecionado o SGBD, feito o modelo físico dos dados, a estimativa do volume de dados, o planejamento de indexação e particionamento.

5. Desenvolvimento e Projeto da Área de Transição: Nesse momento os dados não integrados do ambiente transacional são combinados e transformados em dados corporativos.

São necessários os dados extraídos e transformados para então, serem carregados no Data Mart. Este processo é conhecido como, em inglês, ETL (Extract, Transform and Load), composto pelas etapas de extração, transformação e carga dos dados no Data Mart.

9. Especificação da Aplicação do Usuário Final: É definida qual ferramenta será utilizada para realizar o processamento análitico de fron-ent.

10. Desenvolvimento da Aplicação do Usuário Final: Após as etapas anteriores são realizadas manipulações dos dados além de transformações e produção de gráficos e informação visual.

4. METODOLOGIA

4.1. PLANEJAMENTO DO PROJETO

Escopo: Analisar o perfil dos candidatos que prestaram vestibular da Universidade Federal de Santa Catarina em relação à região e curso escolhido

Justificativa: Qual região busca maior acesso ao vestibular da UFSC, e os cursos mais procurados pelos candidatos. Esta análise pode dizer sobre o futuro das regiões e qual área pode ficar escassa de profissionais.
Exclusões: Foi excluído os candidatos que fizeram o vestibular por experiência.
Riscos: Os dados podem estar inconsistentes ou com erros provenientes da extração do fornecedor.
Fatores Críticos de Sucesso : Responder de forma satisfatória a questões estratégicas e desenvolver um Data Mart corretamente.

4.2. DEFINIÇÃO DOS REQUISITOS DE NEGÓCIO

Com o intuito de obter informações sobre a relação entre os candidatos, região e cursos nas áreas de ensino os seguintes questionamentos foram feitos:

- **a.** Qual ano e cursos de estudantes residentes fora do estado mais foram aprovados?
- **b.** Qual ano e cursos de estudantes residentes no estado mais foram aprovados?
- **c.** Qual a quantidade de inscritos por região?
- **d.** Qual o interesse em cursos de estudantes fora do estado?

4.3. MODELAGEM DIMENSIONAL

Abaixo são apresentados os itens que compõem a modelagem dimensional do Data Mart desenvolvido.
a. Definição do processo a ser modelado: A análise de perfil dos alunos em seus cursos escolhidos, bem como sua região e desempenho dos candidatos nas áreas de ensino.
b. Definição da granularidade: O grão definido foi dos candidatos por curso.
c. Definição das dimensões:

i. dm_candidato: Contém os dados sobre os candidatos e opções escolhidas para o concurso.

ii. dm_evento: Contém as informações dos vestibulares de acordo com cada

ano.

iv. dm_curso: Contém as informações do curso escolhido pelo candidato.

v. dm_regiao: Contém as informações do endereço do candidato.

vi. dm_socioeconomicao: Contém as informações da renda e idade dos candidatos.

d. Definição do fato:

i. fato_desempenho: Contém as informações do candidato por cursos escolhidos, divididos pela sua região de residência, também contém informações sobre a prova e avaliação do candidato..

ii. fato_candidato_notas: Contém as informações do candidato por cursos escolhidos, divididos pela sua região de residência e sua situação socioeconômica.

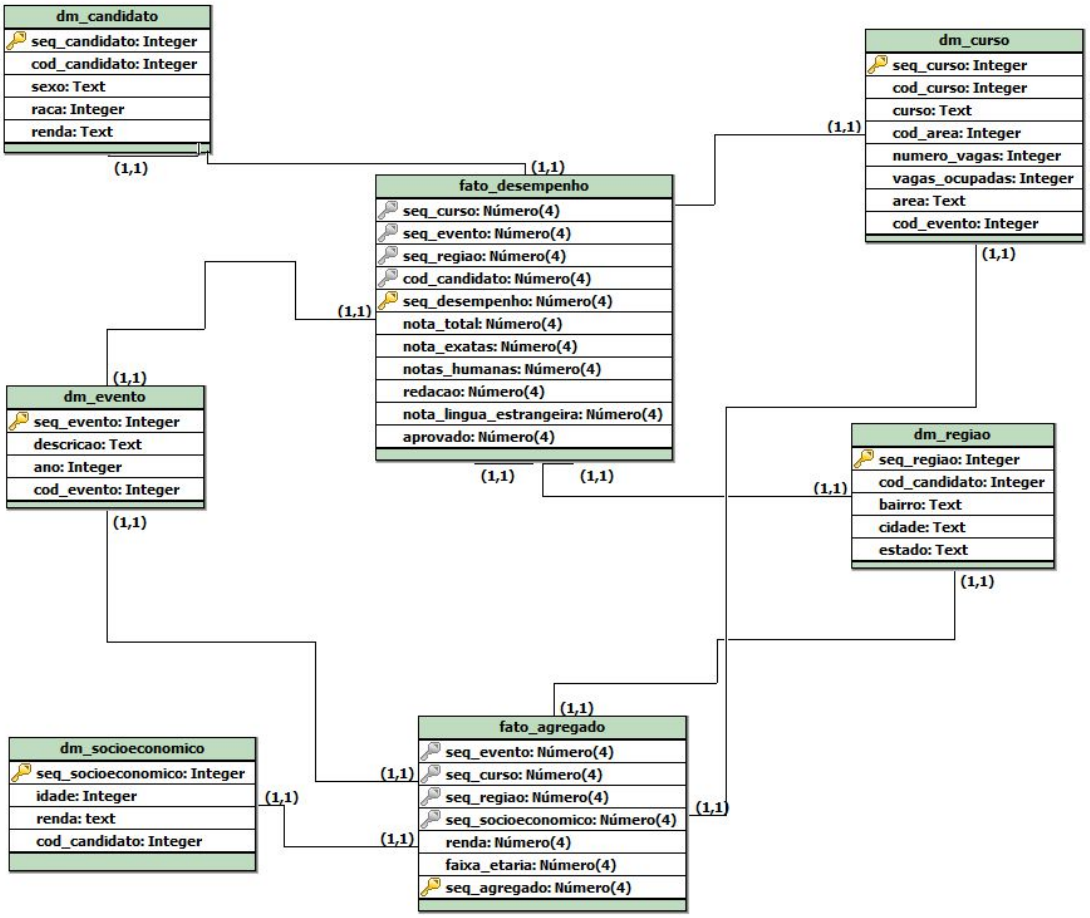


imagem esquema estrela.

4.4. PROJETO FÍSICO

O projeto físico foi desenvolvido conforme apresentado a seguir.

a. SGBD: Para o armazenamento dos dados, foi utilizado o banco de dados Postgres.

b. Modelagem física dos dados: Para a modelagem física dos dados, foi utilizado a ferramenta PGAdmin.

4.5. DESENVOLVIMENTO E PROJETO DA ÁREA DE TRANSIÇÃO

Realizada a conexão com o dump da base de dados disponibilizado pelo Professor. Para acessar estes dados, foi feita uma conexão com DBeaver utilizando docker.

O Kettle serviu para extrair, dar carga aos dados e transformá-los.

Após a exportação os dados foram carregados nas tabelas do Postgress de onde foi feita conexão com o Tableau.

Abaixo, demonstraremos os estágios para preencher as dimensões e os fatos:

a. Dimensão Candidato:

Nesta dimensão são carregados os dados do candidato, foram filtrados os candidatos que fizeram o vestibular por experiência.

b. Dimensão Evento:

Dimensão composta pelos eventos de 2008 a 2012.

c. Dimensão Curso:

Contém todos os cursos do vestibular, foi feito um match entre as tabelas curso e área do curso para obter a descrição de cada curso.

d. Dimensão Região:

Onde são carregadas as regiões de onde vem os candidatos.

e. Dimensão Socioeconomico:

Onde são carregadas as informações socioeconômicas dos candidatos.

e. Fato Desempenho:

Na tabela de fatos “fato_desempenho” é possível observar o resultado de toda a análise sendo feita. Esta é a parte mais normatizada da modelagem, sendo que diversos campos são apenas chaves-estrangeiras para as dimensões. Contém informações sobre as provas e aprovação do candidato

f. Fato Agregado:

Na tabela de fatos “fato_agregado” contém informações sobre a idade e renda dos candidatos, questões estas recebidas da dimensão socioeconomico.

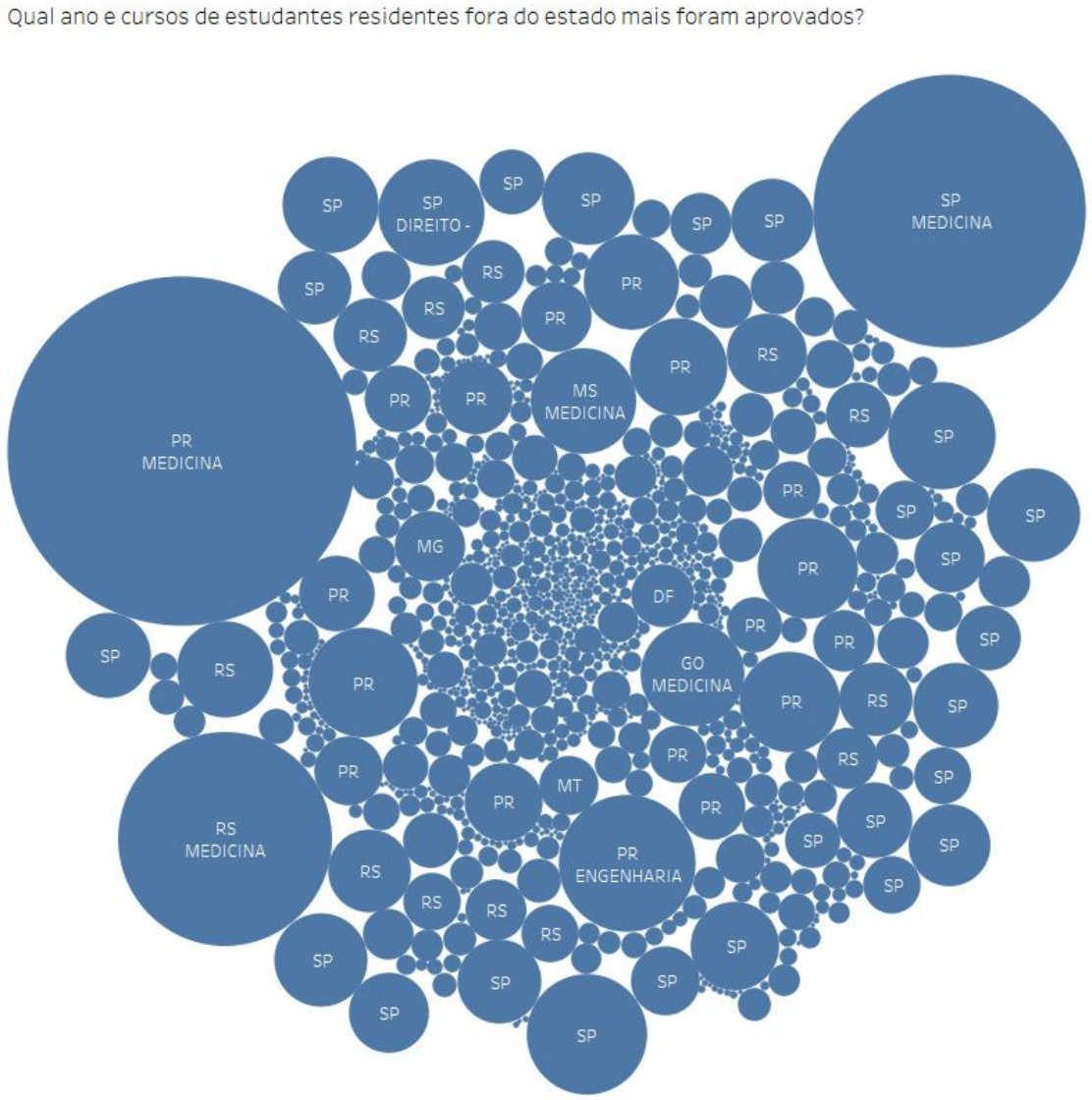
4.6. ESPECIFICAÇÃO DA APLICAÇÃO DO USUÁRIO FINAL

Para o front-end do projeto, onde serão apresentados os gráficos e tabelas para análise, foi utilizado o Tableau.

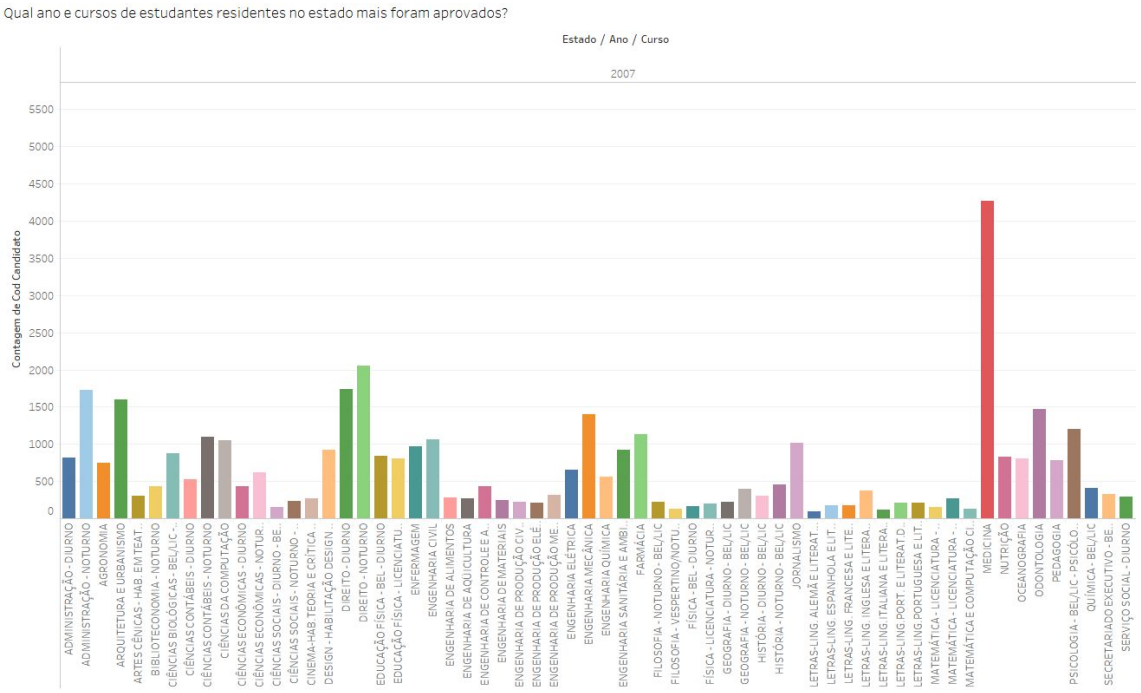
5. RESULTADOS

Os resultados e conclusões das perguntas estratégicas definidas no escopo do projeto obtidos por meio da ferramenta Tableau serão apresentados a seguir.

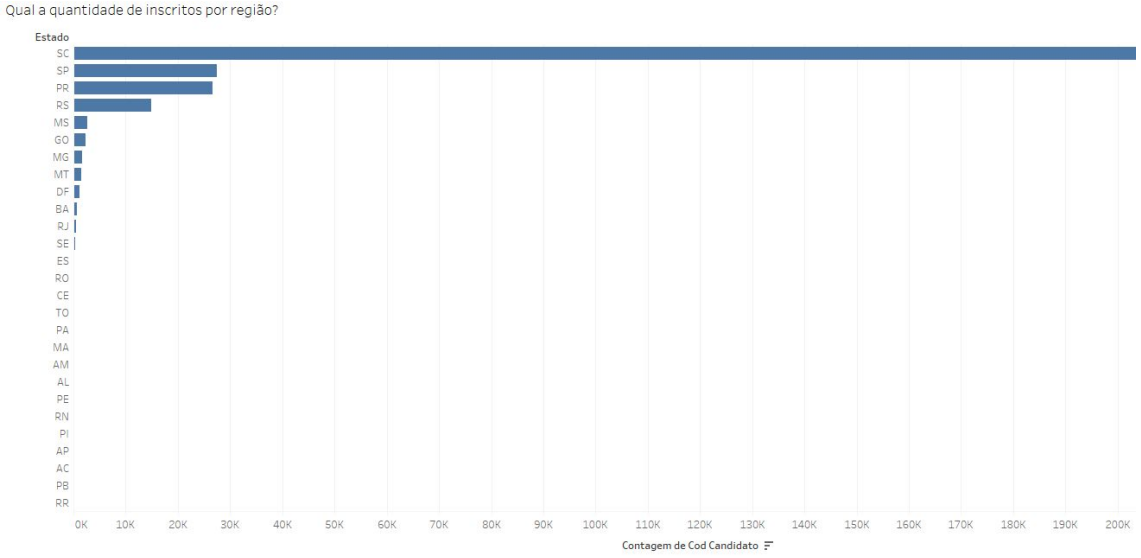
1. Qual ano e cursos de estudantes residentes fora do estado mais foram aprovados?



2. Qual ano e cursos de estudantes residentes no estado mais foram aprovados?



3. Qual a quantidade de inscritos por região?



4. Qual o interesse em cursos de estudantes fora do estado?

6. CONCLUSÃO

Com este artigo, é possível concluir o estudo dos dados dos candidatos ao vestibular UFSC. Os dados são muito importantes para verificar a qualidade da procura e o motivo das procuras pelos cursos da UFSC, com os dois fatos desenvolvidos é possível responder muitas questões socioeconômicas e indicativos sobre os candidatos, infelizmente não foi possível desenvolver grandes quantidades de perguntas, porém com os Data Marts desenvolvidos muitas outras questões podem ser resolvidas.

Para acessar estas informações e reproduzi-las, acesse o endereço https://github.com/thiagodsti/dw-vestibular, e siga os passos descritos no ReadMe.

7. REFERÊNCIAS

[1] Construção de um Data Warehouse em 7 etapas. Disponível em: <https://corporate.canaltech.com.br/tutorial/banco-de-dados/a-construcao-de-um-data-warehouse-em-7-etapas/>. Acesso em: 03 jul. 2017

[2] Construção de um Data Warehouse em 7 etapas. Disponível em: <http://www.oficinadesistemas.com.br/site/03BI/04DataMart/default.html>. Acesso em: 03 jul. 2017

[3] Dump com dados dos candidatos do Vestibular 2008-2012s. Disponível em: <https://www.inf.ufsc.br/~jose.todesco/dw/Vestibular/>. Acesso em: 19 jun. 2017