Análise e desenvolvimento de um Data Mart para investimento de escolas e cursos de pré vestibular em cidades de Santa Catarina

Antonio Homem[1], **Lucas Willrich**[1], **Marina Silva Tavares**[1], **Rhanna Auler**[1], **Thalita Lorenzi**[1]

[1]Departamento de Automação e Sistemas – Universidade Federal de Santa Catarina (UFSC)

amrh07@gmail.com, lucasgermano.w@gmail.com, marinastavares6@gmail.com,

rhannaauler@gmail.com, thalitalorenzi@gmail.com

Abstract. *This document describes the work done for the realization of a Data Mart that aims to find the best municipalities or regions for the investment and implementation of pre-university entrance courses in the state of Santa Catarina, with the objective of supporting the decision making of involved in the process. First, a star scheme was modeled so that it was possible to contain the questions regarding the average of correct answers, the percentage of enrolled, the percentage of those who took the pre-university entrance exams by location. The realization of the Data Mart was possible through the data made available by Coperve for the entrance exams of the Federal University of Santa Catarina between 2008 and 2012, it was also used for the economy of the regions made available at IBGE in 2014. The realization of the back-end was made through Kettle with Pentaho software and Power BI for the front end.*

Resumo. *Este documento descreve o trabalho feito para a realização de um Data Mart que tem como o objetivo encontrar os melhores municípios ou regiões para o investimento e implantação de cursos pré-vestibulares no estado de Santa Catarina, com o objetivo de apoiar a tomada de decisão dos envolvidos no processo. Primeiramente foi modelado um esquema estrela que fosse possível conter as questões referentes a média de acertos, o percentual de inscritos, percentual dos que fizeram curso pré-vestibulares por localização. A realização do Data Mart foi possível através dos dados disponibilizados pela Coperve dos vestibulares da Universidade Federal de Santa Catarina entre anos 2008 a 2012, também foi utilizados referente à economia das regiões disponibilizados no IBGE no ano de 2014. A realização do back-end foi feita através do Kettle com o software Pentaho e o Power BI para o front-end.*

1. Introdução

A concorrência nos vestibulares de faculdades e universidades de todo o país tem aumentado com o passar dos anos. Apesar de alguns cursos possuírem menores relações candidato/vaga, tem sido cada vez mais difícil ingressar em uma boa universidade, também devido ao nível de conhecimento exigido. Esses são os fatores que tem motivado um maior estudo para a implementação de cursos pré-vestibulares, na intenção de aumentar as taxas de aprovação dos candidatos.

Tendo isso em vista, uma rede de cursinhos pré-vestibular da cidade de São Paulo com filiais em diferentes estados do país nos procurou com o interesse de investigar e analisar quais as melhores regiões/cidades em Santa Catarina (SC) para investir em escolas/cursinhos. A empresa pretende ingressar no estado devido a alta demanda de alunos que realizam o vestibular da Coperve, além disso, entre estes muitos são naturais da região.

Dessa forma, será analisado o modelo sócio acadêmico de vestibular da Coperve de 2008 a 2012 com vistas à implementação de um *Data Mart* para suporte e análise dos requisitos levantados. O desenvolvimento do *Data Mart* visa a criação de uma modelagem dimensional até a apresentação das respostas das principais perguntas estratégias em uma ferramenta OLAP (*Online Analytical Processing*).

2. Materiais

Os Materiais utilizados para o estudo das melhores regiões para se investir em escolas/cursinhos em Santa Catarina foram os dados disponibilizados pela Coperve (Comissão Permanente do Vestibular) da UFSC.

Todos os candidatos preenchem um formulário socioeconômico antes de realizar o vestibular, através do qual são coletadas informações sobre alguns aspectos da sua vida escolar, suas condições socioeconômicas e culturais. A prova do vestibular é feita durante 3 dias, contendo questões sobre biologia, física, geografia, história, língua estrangeira, matemática, português, química e redação. As notas são vinculadas aos candidatos no banco de dados e disponibilizadas pela Coperve, juntamente com o formulário socioeconômico e outras informações, como os cursos em que se inscreveram, boletim de desempenho, quantidade de acertos, classificação geral, entre outros. Nessa base de dados também existem informações do vestibular como um todo, tais como número de inscritos, média de acertos e número de vagas.

Além disso, para classificar os municípios e obter uma visão por região foi utilizada a base de dados de 2010 a 2014 do [IBGE], que contém diversos indicadores das cidades, como PIB e divisão por microrregião ou mesorregião. Esses dados foram integrados ao *Data Mart* para complementar a análise.

3. Métodos

A evolução e o crescimento dos dados junto a necessidade constante de gerar informações relevantes sobre o processo de empresas ou o funcionamento de produtos gerados trouxe a necessidade de criação de SAD, sigla para Sistemas de Apoio de Decisão. Os SADs são processamentos analíticos, que são essenciais para a tomada de decisão. O *Data Warehouse* (DW) é um exemplo de tipo de SAD, sendo um repositório de dados integrados, não volátil e variável em relação ao tempo.

O trabalho deste documento propõe a realização de um *Data Mart*, que são pequenas partes de armazenamentos de um subconjunto de dados, organizados para um grupo de investidores interessados no mercado de pré-vestibular. Considerando a teoria *bottom-up* de Ralph Kimball, em que a união de *Data Marts* resultarão em um DW.

Para a extração de dados a serem analisado no *Data Mart* desenvolvido foi utilizado o *software* Pentaho (Kettle), uma ferramenta *open source* de *Extract-Transform-Load* (ETL). Utilizando essa ferramentas disponibilizada, junto aos dados referentes aos vestibulares da UFSC entre os anos 2008 e 2012, foi possível realizar normalização, atualização e gravação de dados no banco de dados a ser utilizado no *Data Mart*. O Kettle foi uma ferramenta de fácil utilização já que essa possui a funcionalidade de *drag and drop* para a formatação e seleção de dados do banco de dados disponível, definindo o *back-end* do projeto.

Para o desenvolvimento do *front-end* do projeto foi utilizado o *software* Power BI, de autoria da Microsoft, em que a sua principal funcionalidade é fornecer visualizações interativas e funcionalidades para a análise de *Business Inteligence*. No projeto em questão, o *software* foi utilizado para gerar os gráficos nos quais é possível identificar tendências e padrões do comportamento dos vestibulares entre os anos selecionados.

4. Metodologia

A etapa de construção de um *Data Mart* para analisar possíveis investimentos na região de Santa Catarina de cursos de pré-vestibular começou com um estudo da proposta do modelo de dados do Vestibular da UFSC, dos anos de 2008 a 2012. A partir do entendimento dos dados obtidos e da proposta de análise a ser realizada, iniciou-se então o processo de modelagem dimensional.

4.1. Modelo Dimensional

O primeiro passo da modelagem dimensional inicia-se na escolha dos processos que devemos modelar. Já sabemos que nossa questão principal é possibilitar uma análise de regiões onde possa vir a ser interessante investir em um curso de pré-vestibular. Os dados obtidos através da Coperve devem então ser modelados e manipulados de forma a obtermos informações relevantes para essas análises como, por exemplo, quantos inscritos teve um determinada localização, qual foi o percentual de aprovados, se os candidatos fizeram pré-vestibular e qual a renda destes candidatos, já que estes tipos de cursos são geralmente pagos.

Em seguida definiu-se o grão do processo de negócio, o segundo passo da modelagem dimensional. Definimos que um registro da tabela de fatos indicaria o desempenho de um candidato em um certo vestibular, o qual ocorre anualmente. Além da tabela fato anatômica, escolhemos também implementar uma segunda tabela de fatos agregada de desempenho por localização, de forma a facilitar as análises no *front end* e já implementar os cálculos necessários na camada de transição.

Já a terceira etapa é a escolha das dimensões e de seus atributos, tendo em vista que estas dimensões serão aplicadas a cada registro da fato. Pensando na nossa proposta de análise e em uma possível posterior expansão deste *Data Mart*, definimos que seria interessante ter as seguintes dimensões: vestibular, vestibulando, curso e cidade. Os atributos de cada uma destas dimensões serão melhor descritos na próxima subseção, onde o processo de ETL é explicado. Nesta subseção também serão apresentados os fatos mensuráveis escolhidos para as tabelas de fatos, sendo que a escolha destes fatos compõe a quarta etapa da modelagem dimensional. O esquema estrela proposto é apresentado na figura 1, onde é possível também ver os atributos e fatos selecionados.

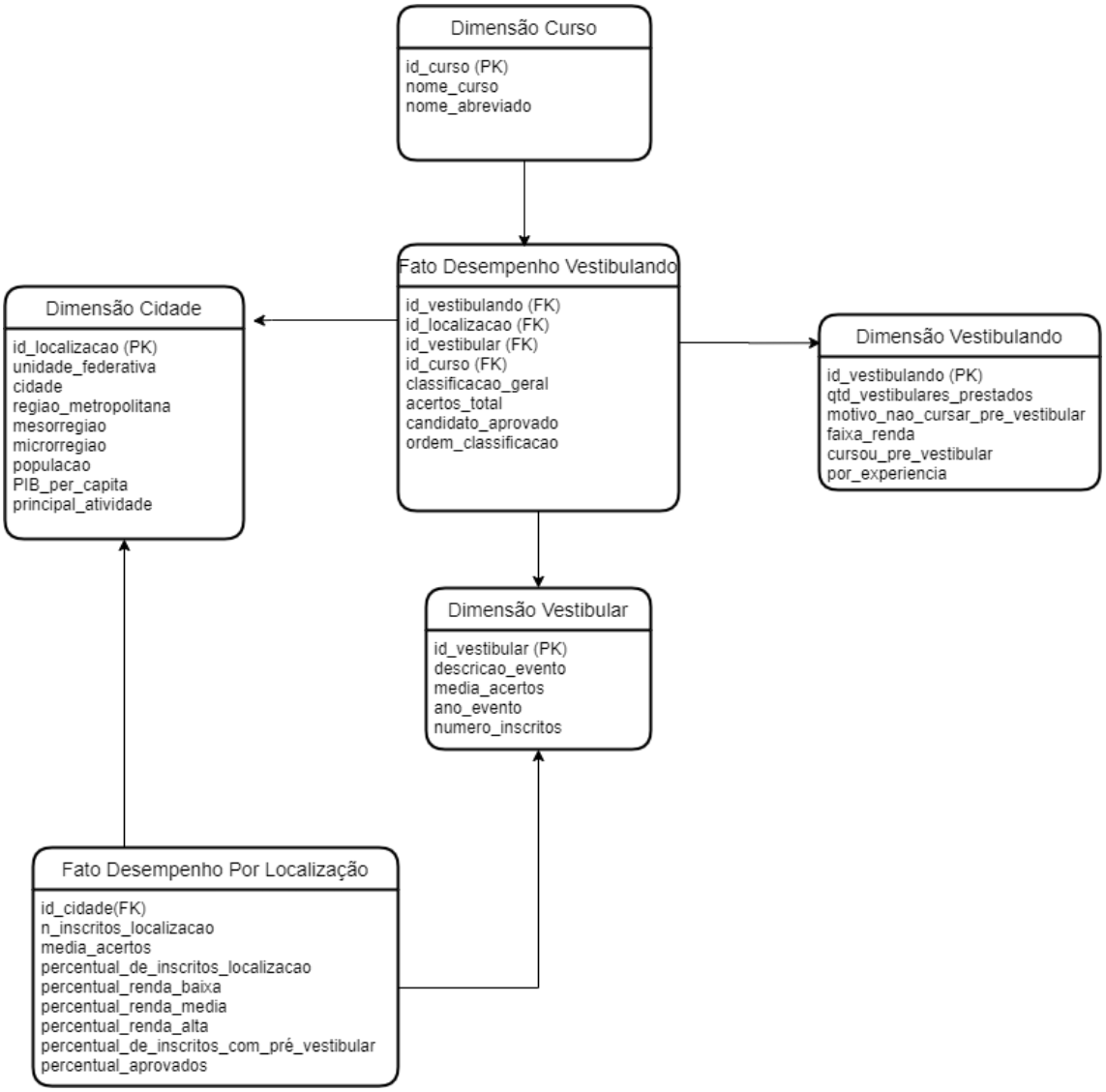

Figura 1. Esquema Estrela

4.2. Processo de ETL

Realizada a modelagem dimensional, parte-se então para a etapa de projeto físico, seguida pelo desenvolvimento e projeto da área de transição. O processo de ETL (extração, transformação e carga) foi realizado utilizando o *software Pentaho Data Integration*.

O processo de ETL da **dimensão Curso** pode ser visto na figura 2. Os dados foram extraídos da tabela curso do transacional, de onde foram selecionados os atributos id_curso, nome_abreviado e nome_curso. Os dados foram então agrupados por id, mantendo o nome do curso que não fosse *null*, e então carregados na tabela de dimensão Curso.

Figura 2. ETL da Dimensão Curso

O processo da **dimensão Vestibular** também foi bastante simples, podendo ser visto na figura 3. Os dados foram extraídos da tabela evento do transacional, modificando o campo do id e então carregando para a dimensão Vestibular.

Figura 3. ETL da Dimensão Vestibular

Já os dados da **dimensão Vestibulando** foram extraídos da tabela do transacional Candidato, como pode ser visto na figura 4. O campo de grade sócio econômica foi tratado de forma a pegar as partes referentes as respostas de se o candidato frequentou pré-vestibular, o motivo por não ter feito esse cursinho, e a faixa de renda do candidato. A partir destas respostas foram criados subgrupos, como no caso da faixa de renda, esta foi dividida entre: baixa, baixa/média, média, média/alta e alta. Já a resposta do pré vestibular foi transformada em apenas ”sim”e ”não”. Após a transformação das respostas, manteve-se apenas os campos desejados na tabela de dimensão, adicionado o id desta e então feita a carga para a tabela de dimensão Vestibulando.

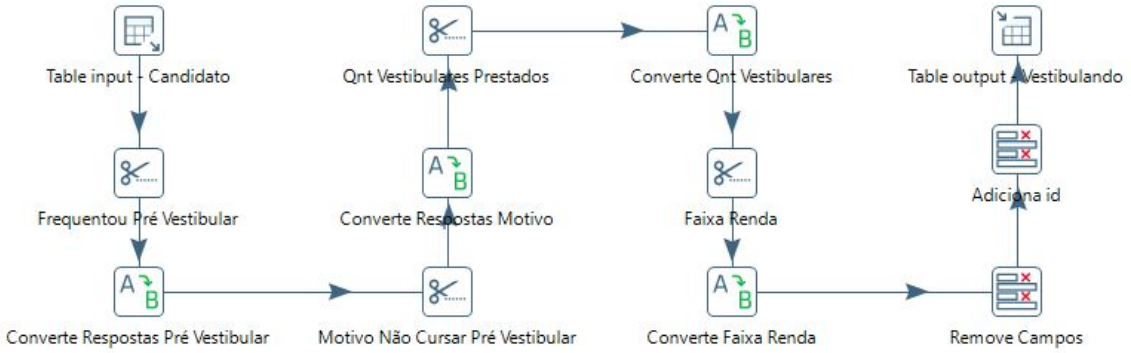

Figura 4. ETL da Dimensão Vestibulando

Para incrementar os dados da **dimensão Cidade**, foram obtidos dados referentes aos municípios brasileiros [IBGE], como a região metropolitana ao qual fazem parte, o PIB do município, mesorregião, microrregião, e principal atividade econômica. Os nomes das cidades e unidades federativas da tabela cidade e da base de dados do IBGE foram então tratados, já que haviam discrepâncias como letras maiúsculas e minúsculas,

com acento e sem, e distinções como ”STO”e ”SANTO”. As cidades que possuem essas diferenças foram então unificadas após o tratamento das strings. A partir do nome da cidade e da unidade federativa foi calculado o id da localização. Após a transformação dos dados, a base de dados foi unida com os dados da tabela Cidade do Transacional, e os dados foram então carregados para a dimensão Cidade.

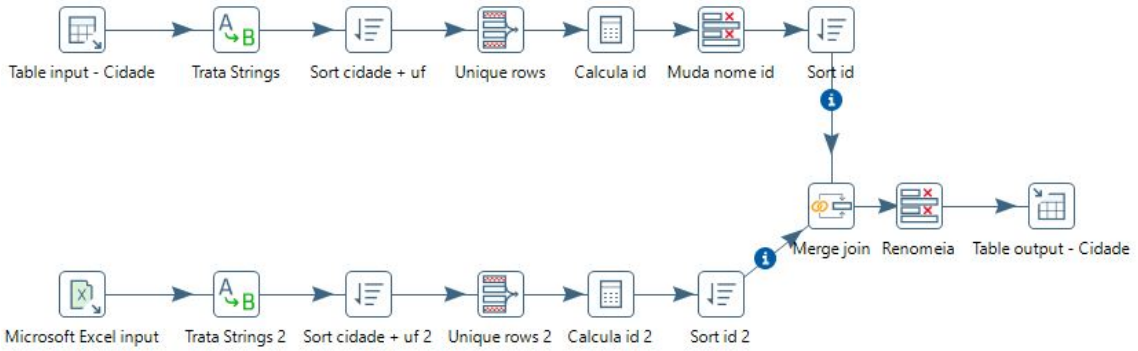

Figura 5. ETL da Dimensão Cidade

O processo de construção da tabela **Fato Desempenho por Candidato** pode ser visto na figura 6. Inicialmente, os dados das tabelas Candidato e Candidato Aprovado do Transacional são obtidos, de forma a criar um campo que indica se o candidato foi aprovado ou não. Os dados são então unidos aos da dimensão Vestibulando, através do id do candidato. Os campos cidade e unidade federativa da tabela Candidato são então unidos para formar o id cidade, o qual é usado para unir os dados com os o da dimensão cidade. Em seguida, o id do evento é usado para unir os dados da transformação com os da dimensão vestibular. Já para fazer a relação com a dimensão curso é necessário obter o id do candidato e relacioná-lo com o curso através da tabela Opção Candidato do Transacional, e em seguida uni-lá aos dados tratados. Após fazer a relação com cada tabela dimensão, os dados são então carregados para a tabela Fato Desempenho por Candidato.

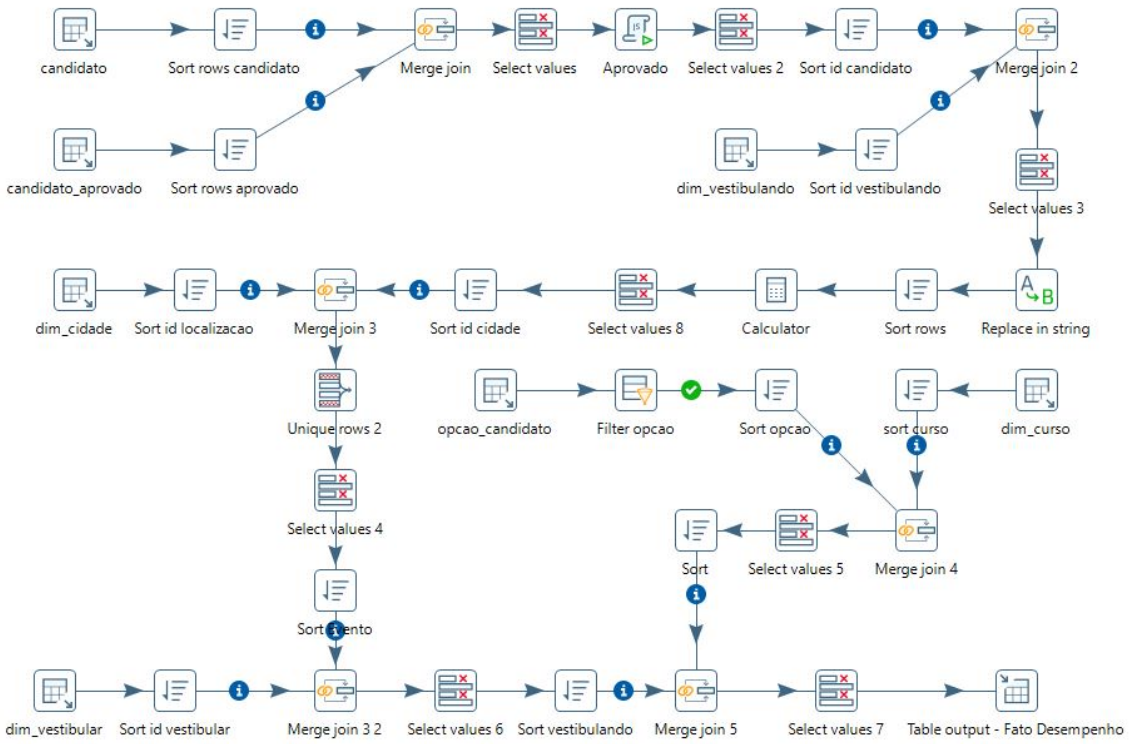

Figura 6. ETL da Fato Desempenho por Candidato

A tabela **Fato Desempenho por Localização** possui um processo bastante similar ao da tabela fato anterior, como pode ser visto na figura 7. No entanto, acaba não tendo relação com o curso e a relação com o vestibulando é eliminada após pegar os dados necessários para a agregação. As colunas referentes à grade sócio econômica da dimensão Vestibulando são utilizadas para calcular novos campos numéricos que indicam se o candidato fez pré vestibular e se possui renda baixa, média ou alta. Um *step group by* foi então usado para agrupar os dados por cidade e vestibular. Em seguida, usou-se o código *javascript* da figura 8 para calcular os campos de percentuais da tabela fato Desempenho por Localização, a qual foi carregada após a definição dos campos.

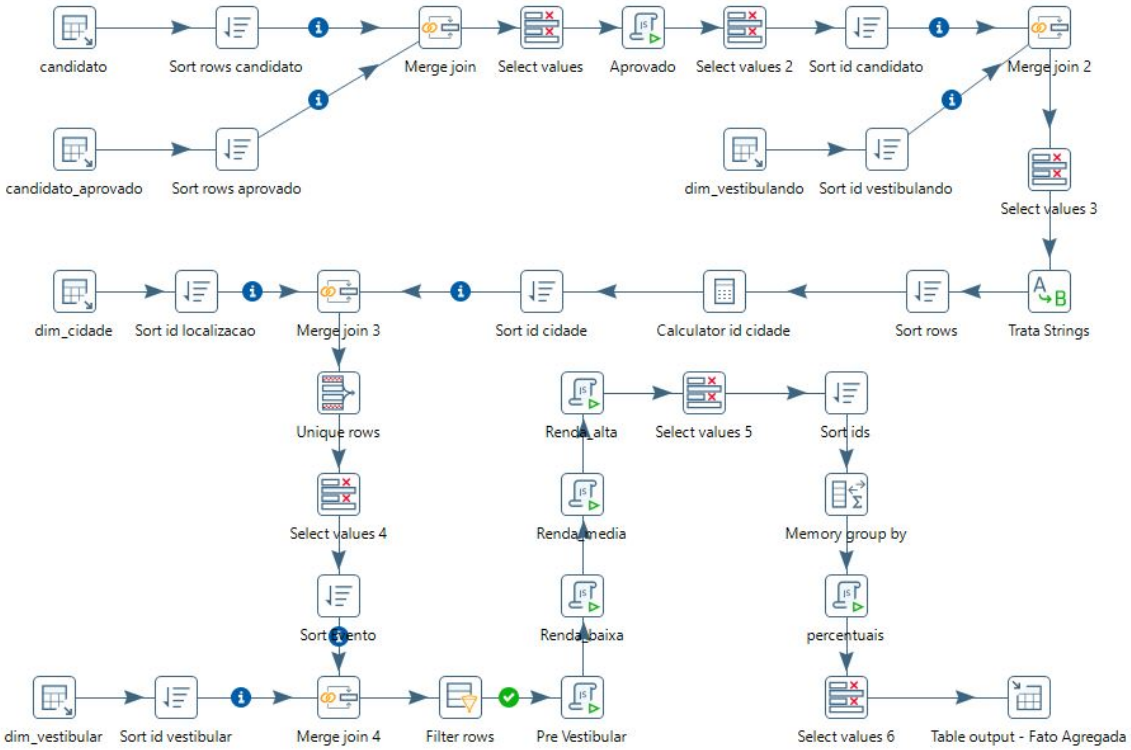

Figura 7. ETL da Fato Desempenho por Localização

```
Script 1
var percentual_inscritos_localizacao
percentual_inscritos_localizacao = n_inscritos_localizacao / numero_inscritos
var percentual_renda_baixa
percentual_renda_baixa = renda_baixa/n_inscritos_localizacao

var percentual_renda_media
percentual_renda_media = renda_media/n_inscritos_localizacao

var percentual_renda_alta
percentual_renda_alta = renda_alta/n_inscritos_localizacao

var percentual_de_inscritos_com_pré_vestibular
percentual_de_inscritos_com_pré_vestibular = fez_pre/n_inscritos_localizacao

var percentual_aprovados
percentual_aprovados = aprovado/n_inscritos_localizacao
```

Figura 8. Código *JavaScript* usado para cálculo dos percentuais

O *job* implementado para a realização de todas as transformações desenvolvidas no *Pentaho* pode ser visto na figura 9.

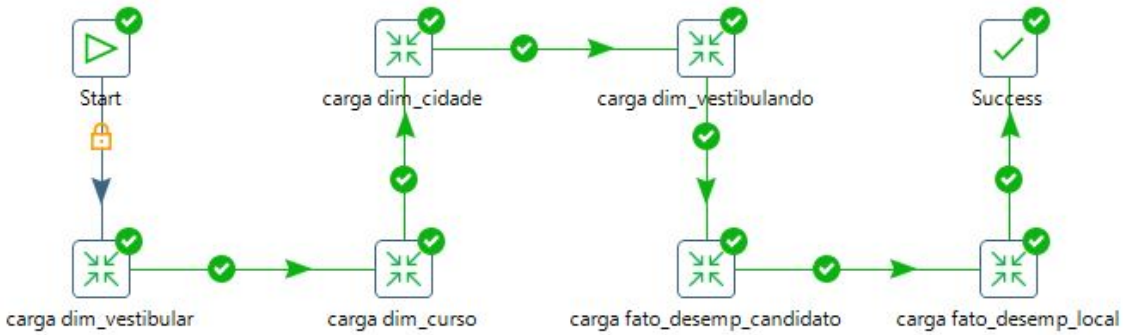
Start
carga dim_cidade
carga dim_vestibulando
Success
carga dim_vestibular
carga dim_curso
carga fato_desemp_candidato
carga fato_desemp_local

Figura 9. ***Job***

5. Resultados

Para realizar a análise de qual é a região mais favorável de Santa Catarina pra se investir em um cursinho pré-vestibular foram geradas diversas visualizações através do *software* Power BI com objetivo de responder algumas perguntas importantes.

A primeira questão respondida foi:

- Qual é a média de acertos por localização?

Para isso foram geradas duas visualizações, figura 10 e figura 11, as quais apresentam graficamente as cidades e regiões, respectivamente, com maiores médias de acertos no estado. Concluímos dessas imagens que há uma diferença significativa entre as médias de acertos das diferentes localizações.

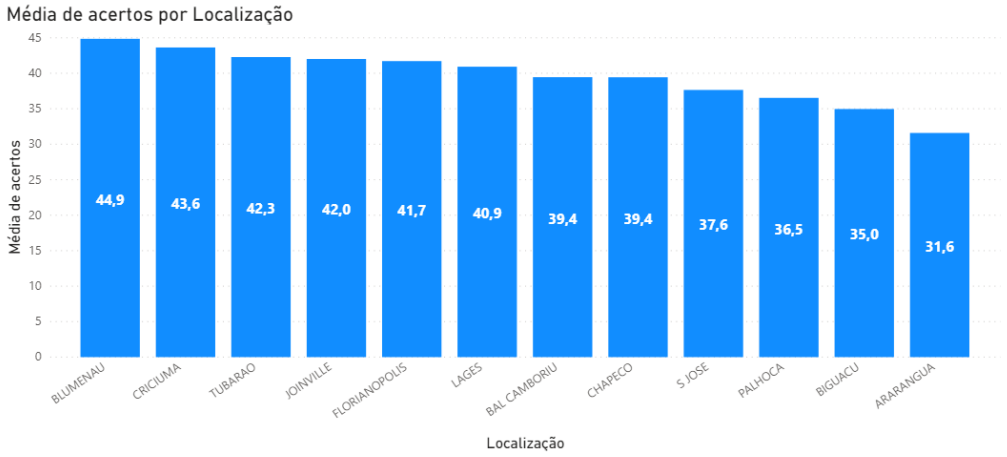

Figura 10. Média de acertos por Localização

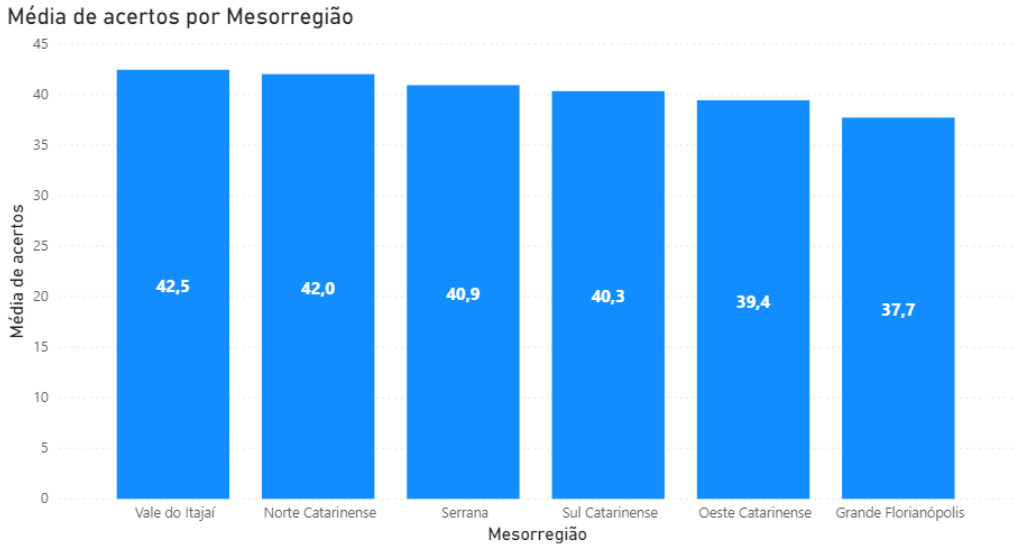

Figura 11. Média de acertos por Mesorregião

Em seguida respondeu-se a questão:

- Qual é o percentual de inscritos por localização em relação ao total do vestibular?

Inicialmente observando a distribuição de inscritos por região do estado na figura 12 percebe-se que a região com maior volume de inscritos é a Grande Florianópolis, seguida do Sul Catarinense. Para entender melhor essa distribuição aprofundou-se a visualização realizando *drill-down* da região, chegando em um nível maior de detalhamento na visualização dos inscritos por localização, apresentada na figura 13, onde é possível avaliar quais cidades possuem maior número de candidatos.

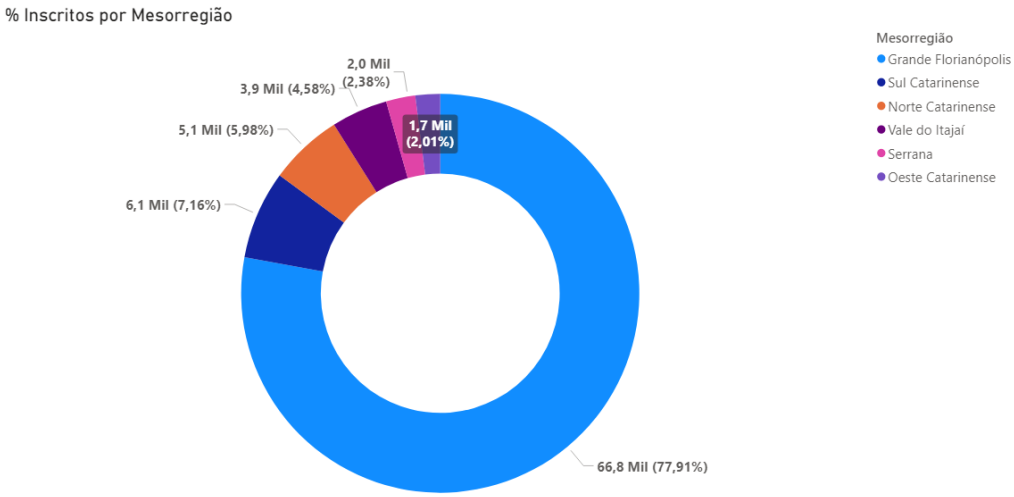

Figura 12. Inscritos (%) por Mesorregião

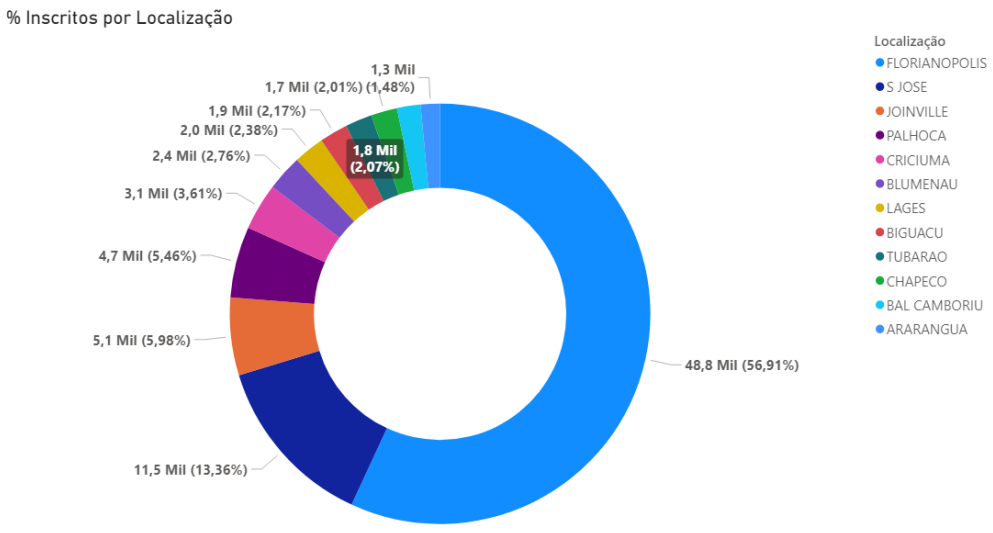

Figura 13. Inscritos (%) por Localização

A seguinte questão respondida foi:

- Qual é a renda média por região?

A figura 14 exibe o gráfico gerado para responder esse questionamento. A partir desse entendimento torna-se mais claro quais são as regiões com potencial econômico para receber o possível investimento. O Norte Catarinense, o Vale do Itajaí e a região Serrana são as áreas mais propícios nesse âmbito.

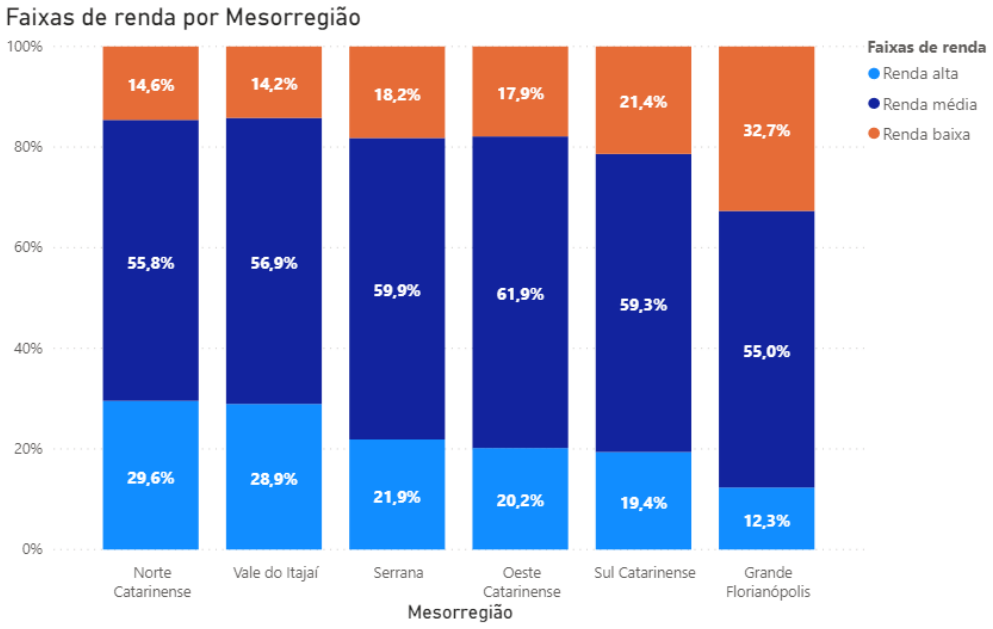

Figura 14. Distribuição das Faixas de Renda por Mesorregião

Para compreender quais locais apresentam maior volume de concorrentes para o negócio foi proposto o seguinte questionamento:

- Qual é o percentual de inscritos que fizeram pré-vestibular por localização?

Através do gráfico de barras apresentado na figura 15 é possível identificar as localizações onde a proporção de candidatos que realizaram cursinho pré-vestibular é maior. Dessa forma, descobriu-se que existe uma atuação forte de outras redes de cursinhos em Chapecó, Joinville, Florianópolis e Balneário Camboriú.

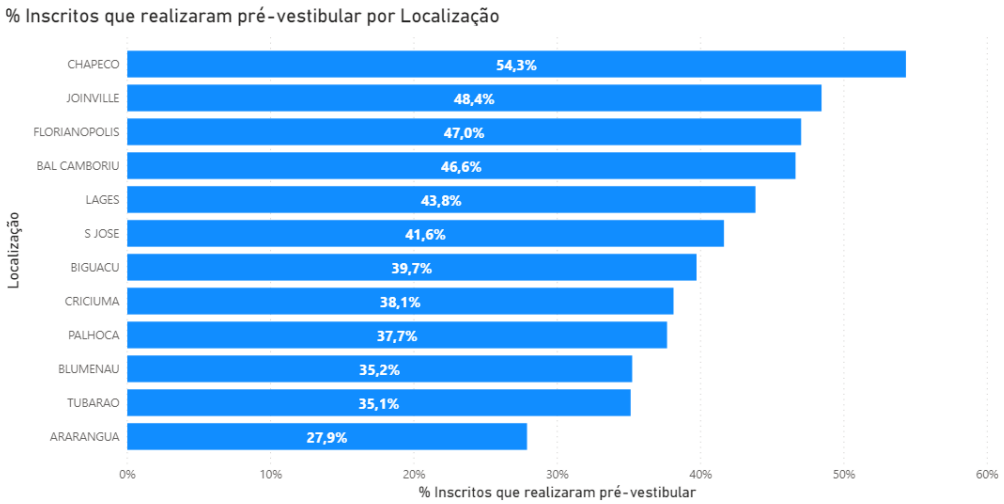

Figura 15. Inscritos (%) que realizaram Cursinho pré-vestibular por Mesorregião

A última questão respondida foi:

- Qual é o volume médio de aprovados por região?

Novamente, a análise foi iniciada através de uma visão macro, onde foram observadas as médias de aprovados por região do estado. Em seguida, foi realizado o *drill-down* para o nível de cidade, onde foi analisado o desempenho dos candidatos com maior detalhamento, como exposto na figura 17. Através dessa investigação percebeu-se que embora as cidades de Balneário Camboriú e Chapecó apresentem um cultura local de realização de cursinhos, como exposto na resposta do questionamento anterior, o desempenho dos candidatos não está entre os melhores observados, indicando uma possibilidade interessante para investimento. Contudo, é necessário realizar análises mais detalhadas posteriormente para compreender se de fato esses locais são adequados para o projeto.

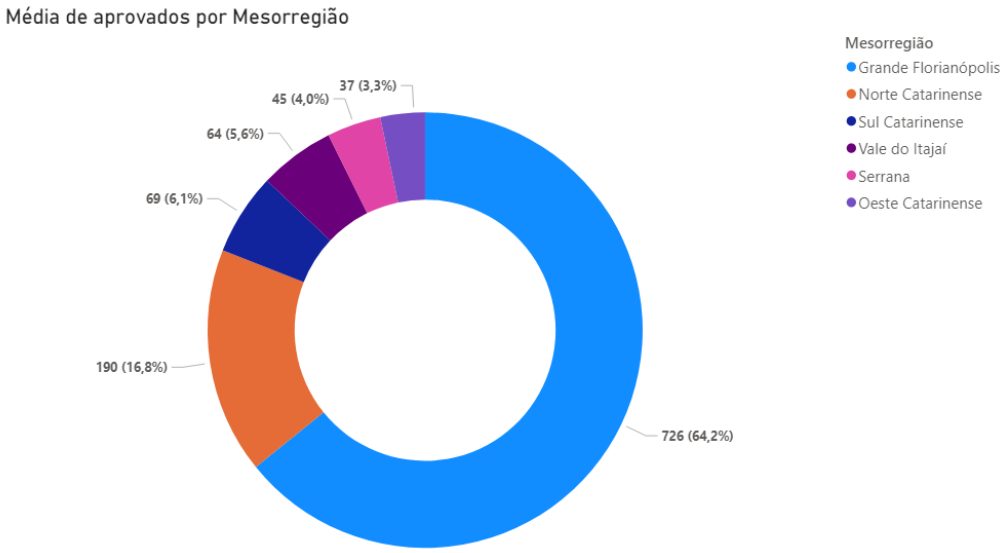

Figura 16. Volume médio de aprovados por Mesorregião

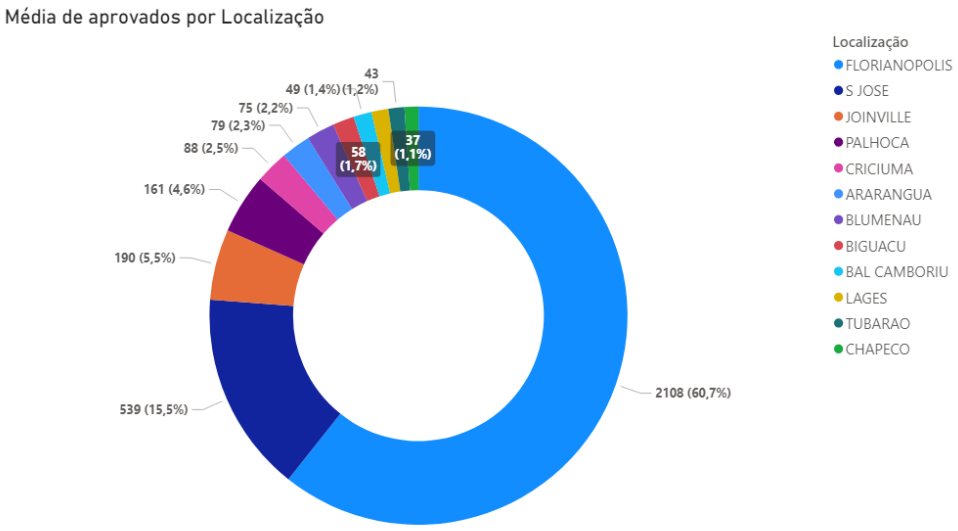

Figura 17. Volume médio de aprovados por Localização

6. Conclusões e Trabalhos Futuros

A partir das respostas obtidas para as perguntas estratégicas, observamos os seguintes pontos:

- Florianópolis possui a maior quantidade de aprovados por região, no entanto, isso se deve a grande quantidade de inscritos de lá. Ao observar a média de acertos nas provas, a Grande Florianópolis possui o pior desempenho, mostrando que existe mercado para a implantação de novos cursinhos, ainda que porcentagem de pessoas com renda baixa seja uma das maiores do estado.
- O Sul Catarinense se destaca por ter uma das maiores rendas do estado, além de concentrar a segunda maior porcentagem de inscritos do vestibular.
- Já o Vale do Itajaí, além de ter uma renda e quantidade de inscritos em destaque, apresenta a maior média de acertos do vestibular. Porém esse valor ainda é menor que a maioria das notas de corte dos cursos da UFSC. Nessa região a porcentagem de inscritos que realizaram curso pré-vestibular também é baixa.

Devido a todos os fatores elencados acima, definiu-se que as regiões mais atrativas para a implantação das filiais da rede de curso/colégio pré-vestibular de Santa Catarina são: A Grande Florianópolis, Vale do Itajaí e Sul Catarinense. Com destaque para as cidades de Criciúma, Blumenau, Balneário Camboriú e Florianópolis, que possuem melhor risco/retorno.

Por último vale destacar que o desenvolvimento de um *Data Mart* utilizando as ferramentas propostas foi essencial para a compreensão e consolidação de teorias aprendidas durante a disciplina de DataWarehouse. Apesar da dificuldade da instalação do *software* Pentaho, a manipulação dos dados para a confecção do *back-end* do projeto junto aos dados oferecidos foram suficientes para entender e oferecer respostas conclusivas sobre o problema tratado.

Referências

IBGE. Pib municípios - base de dados 2010 - 2014. ftp://ftp.ibge.gov.br/Pib_Municipios/2014/base/base_de_dados_2010_2014.xls.