Datar Mart para características e ambientes de vestibulandos

André Almeida Pfeiffer[1]

[1]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC) – Florianópolis, Santa Catarina - Brasil

{pfeiffer}@inf.ufsc.br

Abstract. *Aiming the implementation of a Data Mart based on Comissão Permanente do Vestibular (COPERVE) datas and to find personal and environmental characteristics of university entrance exam applicants, who exceled on success during UFSC 2008 to 2012 entrance exams, took place this presentation who resulted on analisis of the profiles who are most likely to succeed.*

Resumo. *Visando a implantação de um Data Mart baseado em dados fornecidos pela Comissão Permanente do Vestibular (COPERVE) e com foco em determinar as características pessoais e ambientais de quem atingiu excelência nos vestibulares UFSC 2008 a 2012 desenvolveu-se o trabalho apresentado pelo presente artigo que resultou em uma análise do perfil propicio à aprovação.*

1. Introdução

O presente artigo descreve o trabalho de implementação de um Data Mart. Para tanto, é necessário entender que um Data Mart visa dar suporte à análise e tomada de decisões e que para tal objetivo o cruzamento de dados é uma importante ferramenta de geração de indicadores que serão usados para tomada de decisão.

O objetivo central do Data Mart desenvolvido é compreender o perfil dos estudantes que tiveram melhor desempenho e assim determinar algumas características que podem ser determinantes no sucesso da obtenção de aprovação no vestibular de modo que gestores possam definir políticas ou cursos de ações para melhorar o quadro da educação fundamental e média onde mais precisar de atenção.

Assim sendo será apresentado nesse artigo alguns conceitos da área de Data Warehouse por meio da descrição da implementação. Serão abordados Métodos, Metodologia, Escopo bem como Materiais e os Resultados da presente obra.

2. Materiais

A SETIC é a entidade da UFSC responsável por armazenamento digital de informações. Apesar dos dados serem fornecidos pela SETIC, cabe se ressaltar que a Comissão Permanente do Vestibular (COPERVE) quem autoriza a liberação dos dados. É a COPERVE quem elabora os questionários e provas do vestibular bem como guias. Dos produtos da COPERVE o questionário socioeconômico faz-se uma ferramenta poderosa para identificação dos perfis desejados. Com o questionário em mãos verifica-se a

possibilidade do desenvolvimento do trabalho uma vez que o mesmo levanta dados inerentes ao ambiente e características pessoais do candidato.

3. Métodos

Definição do escopo: Traçar o perfil dos 500 melhor classificados vestibulandos na base de dados fornecida.

Justificativa: identificar características que levam ao alto desempenho de modo a definir um ambiente e perfil mais favorável ao sucesso

exclusões do escopo:
Definir o sucesso da média geral;
Definir indicadores de fracasso;
Exclui-se análises referentes às cidades;

fatores críticos de sucesso: Capacidade de determinar perfil de sucesso.
Responder as perguntas:

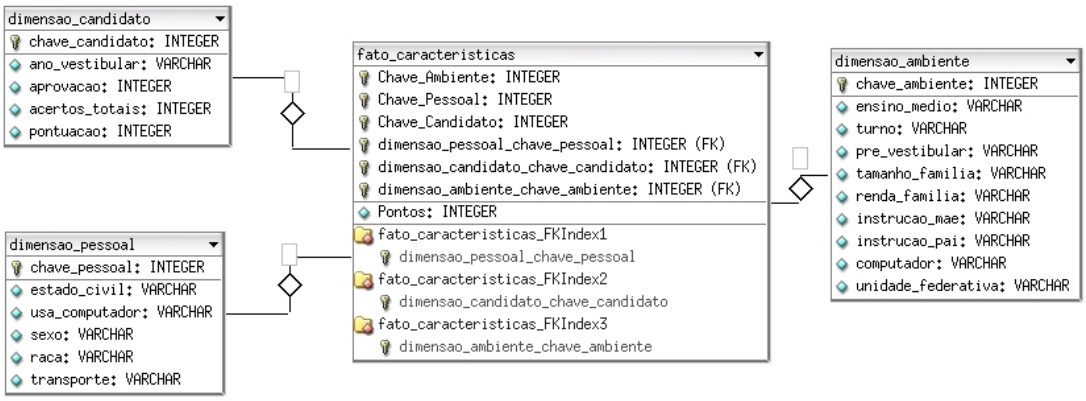
1 – Qual o tipo de ensino garante maior aproveitamento, por unidade federativa;
2 – Qual a raça possui melhor aproveitamento;
3 – O sexo da candidato influencia no resultado;
4 – Quanto a influencia do nível educacional dos pais reflete sobre o desempenho do candidato;

riscos: abstração do mundo por meio dos dados ser insuficiente para dar uma resposta precisa.

4. Metodologia

4.1. Dimensões

Como pré-requisito para o desenvolvimento do presente trabalho foi necessário se estabelecer as dimensões que seriam utilizadas para se responder as perguntas estabelecidas como fator crítico de sucesso. Com base em modelo dos dados as seguintes dimensões foram arquitetadas:



4.2. Identificação do projeto

A identificação do projeto é uma abordagem opcional no desenvolvimento de um Data Mart. Para fim de ilustração foi adotado o seguinte nome para o Data Mart: top of the point .

Em adição, tomou-se como padrão de nomenclatura o “snake_case”.

4.3. Desenvolvimento do plano

ETL ou Extract, Transform, Load são etapas de obtenção de dados e tratamentos do mesmo. Para se realizar o processo de ETL fora utilizado a ferramenta Kettle do aplicativo Pentaho. Essa ferramenta permite definir fluxos dos dados e seus tratamentos bem como transformações e cargas.

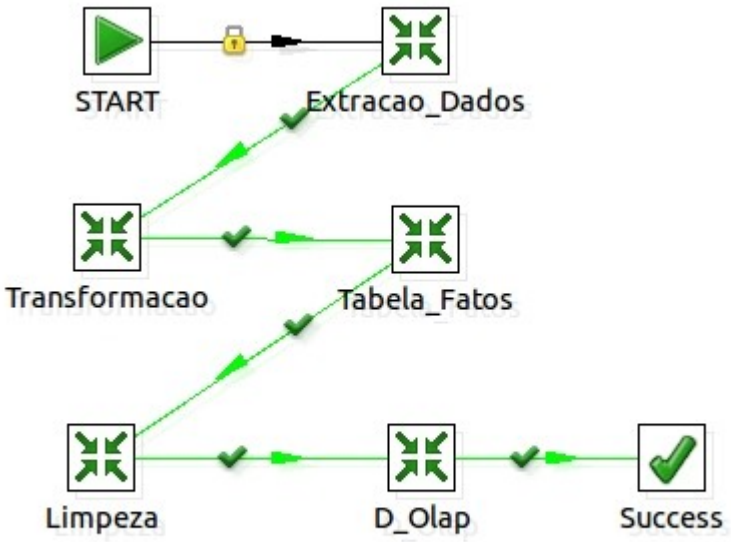


Figura 1: Job com os passos realizados no processo de ETL

Na Figura 1 é possível se verificar o fluxo de uma maneira geral e como foi subdividido em passos. O primeiro extrai informações importante para resolução do problema. Em seguida esses dados passam pelo estágio de transformação onde dados codificados são transformados em dados que possam ser utilizados por humanos. Após, é iniciado a população da tabela de fatos com todas as chaves e uma pontuação atribuída para aquele conjunto (candidato + ambiente + pessoal). Em seguida dados desnecessários para o estágio final são excluídos. Por último uma tabela é populada para se estabelecer o desktop olap (D_Olap).

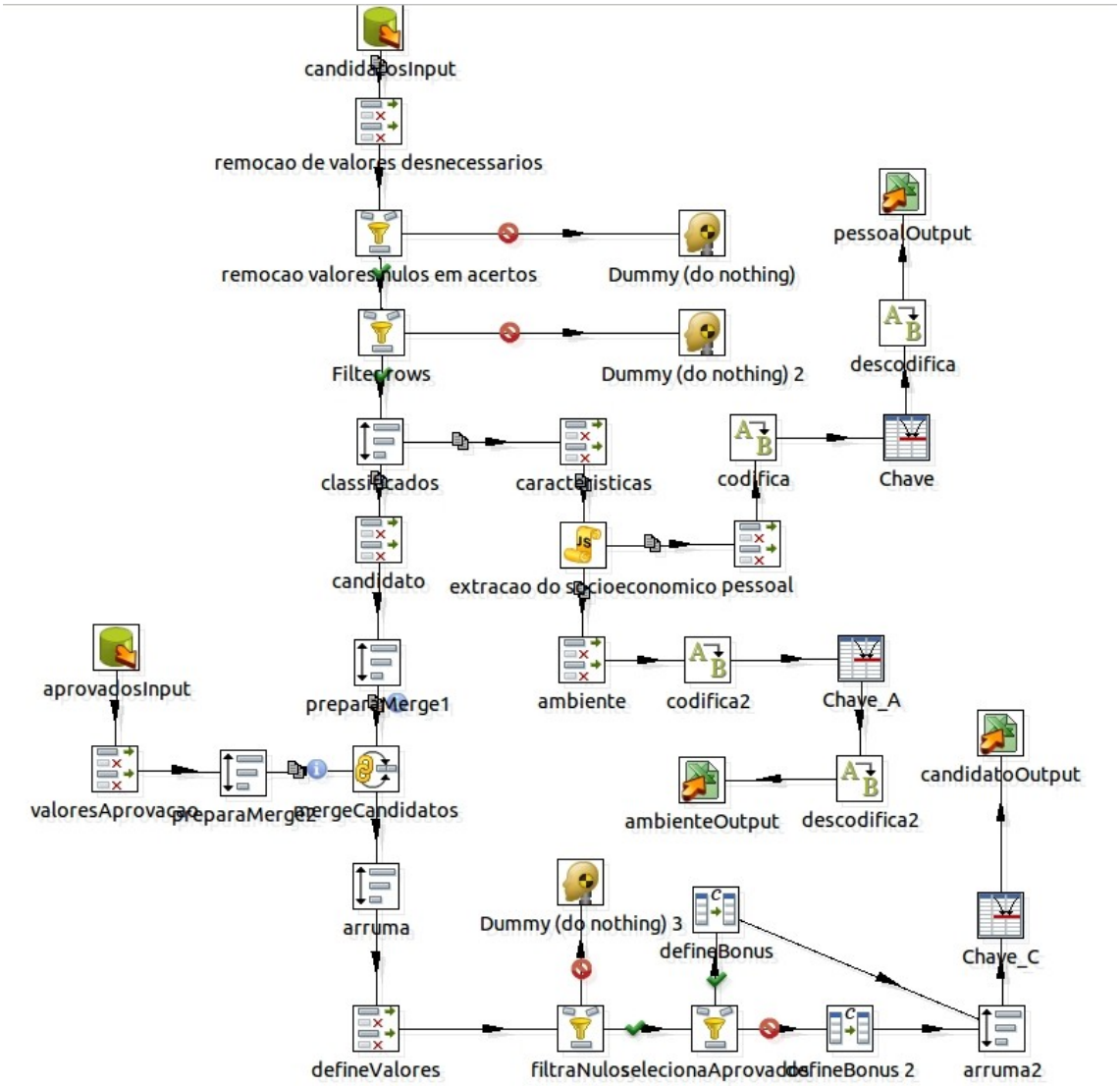


Figura 2: processo de extração das dimensões

Na Figura 2 é possível se ver o processo de extração das dimensões candidato, ambiente e pessoal. Os dados são carregados do banco de dados e são excluídos as informações que fogem do escopo do presente trabalho. Em seguida dois filtros são aplicados para se remover valores nulos em acertos e classificação geral para diminuir os dados que serão trabalhados e evitar problemas com estes valores. Os dados são ordenados e divididos. A porção de características possui os valores do questionário todos codificados em uma única coluna. Esses valores são separados em diversas colunas por meio do script que se segue:

```
var estado_civil = gradesocio.getString().substr(0,1);
var estado_origem = gradesocio.getString().substr(1,2);
var ensino_medio = gradesocio.getString().substr(9,1);
var turno = gradesocio.getString().substr(10,1);
var pre = gradesocio.getString().substr(12,1);
var familia = gradesocio.getString().substr(18,1);
var salario_familia = gradesocio.getString().substr(19,1);
var instrucao_pai = gradesocio.getString().substr(20,1);
var instrucao_mae = gradesocio.getString().substr(21,1);
var computador_residencia = gradesocio.getString().substr(27,1);
var usa_computador = gradesocio.getString().substr(28,1);
var transporte = gradesocio.getString().substr(29,1);
```

Após, são divididos nas dimensões pessoal e ambiente onde são criadas chaves baseadas no conjunto de características que neste ponto ainda estão codificadas. Os candidatos são mergidos aos candidatos classificados pelas suas id de candidato e de evento e logo após recebem tratamentos para valores nulos novamente para prevenir erros. Um valor bonus é definido da seguinte maneira: caso o candidato tenha sido aprovado no vestibular, independente da sua classificação geral, é atribuído um bonus de 25% (1,25). A chave do candidato passa a ser o id do candidato concatenado com o id do evento.

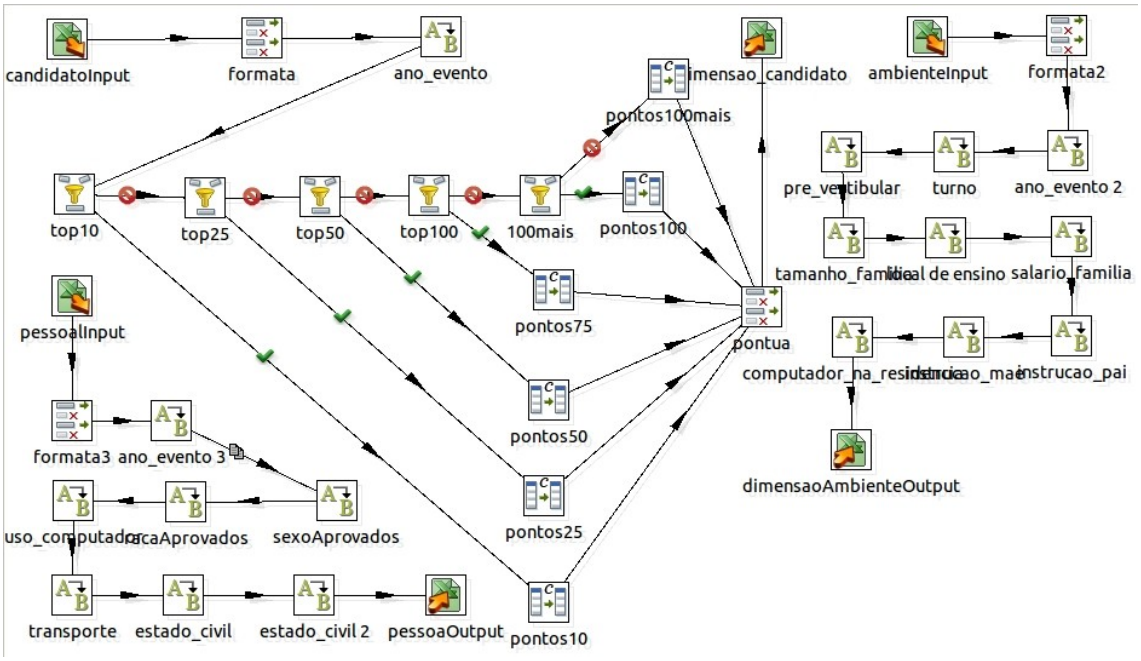


Figura 3: transformação dos dados

Na transformação os dados são formatados para string de modo a se evitar a perda de formatação dos números, causada pelo kettle e que mesmo com as opções de formatação da ferramenta acabam por perder-se. À dimensão de candidato é aplicado uma série de filtros para pontuar a classificação. Aos 10 primeiros são atribuídos 100 pontos, aos 25 primeiros (de décimo primeiro ao décimo quinto) são atribuídos 75 pontos e assim se repete com os 50, 75, 100 e mais de 100 da classificação geral com respectivamente 65,

60, 55 e 50 pontos. Às dimensões ambiente e pessoal segue-se a transformação dos códigos em valores legíveis.

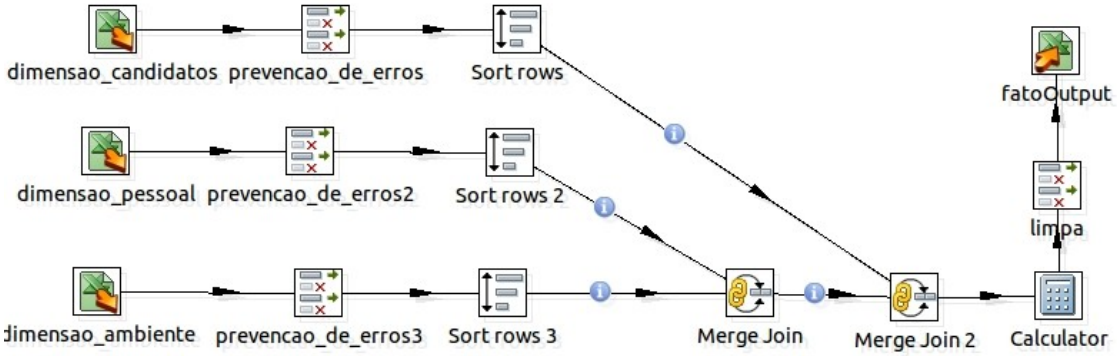


Figura 4: carga dos fatos.

Neste passo os dados são organizados e as chaves de cada um são selecionadas com base no id do candidato e do evento. Uma pontuação é calculadas com base no total de acertos multiplicado pelo bônus somado à pontuação pela posição geral calculada no passo de transformação dos dados.

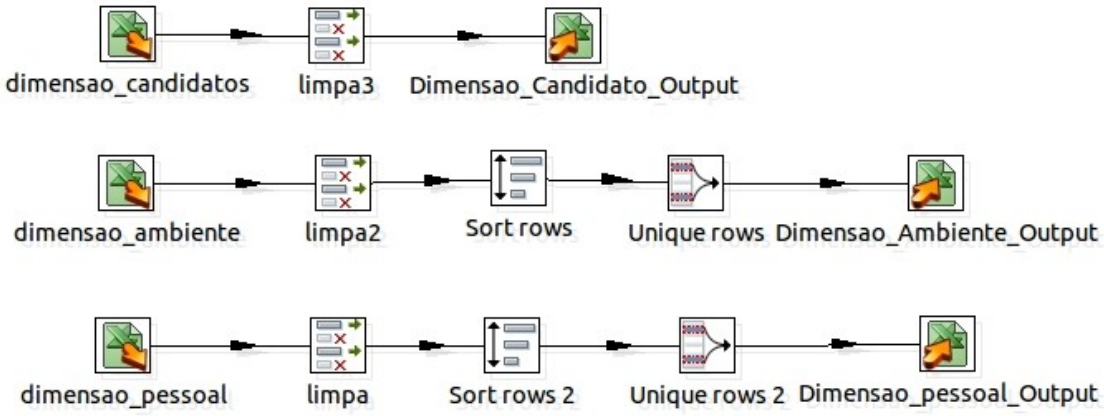


Figura 5: Limpeza das dimensões.

Ao candidato é excluído os ids de candidato, que neste ponto tornaram-se desnecessários. Os ids de evento são traduzidos para o ano de realização do vestibular. Às dimensões ambiente e pessoal são excluídos todos os campos desnecessários e que não fazem parte destas dimensões. São organizadas pelas chaves da respectiva dimensão e valores repetidos são excluídos.

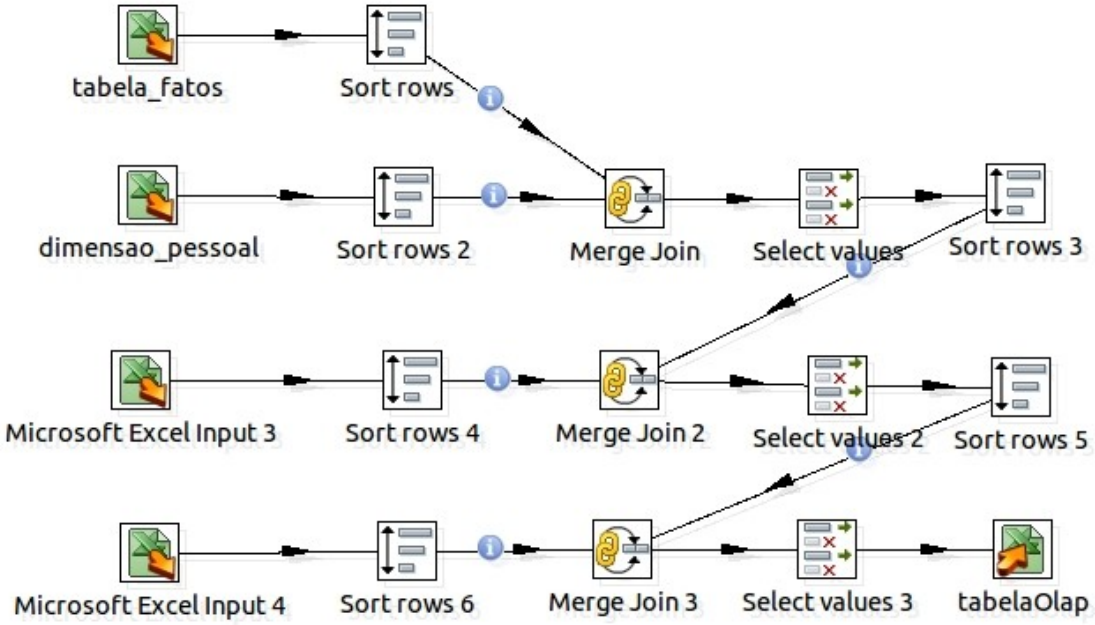


Figura 6: extração da tabela Olap.

Neste ponto é possível se fazer consultas ao data mart bem como armazená-lo em algum SGBD. Para efeito ilustrativo e de legíbilidade esta etapa foi feita no kettle. Para qualquer consulta basta buscar pela combinação das chaves da tabela de fatos. No exemplo todos os dados são agrupados para se possibilitar o desenvolvimento da tabela dinâmica.

4.4. Olap

Para o processo de OLAP fora utilizado as tabelas dinâmicas do LibreOffice com pivoteamento. Como os dados já encontravam-se em uma tabela excel, bastou apenas a criação do pivoteamento dos dados inerentes ao aspecto ambiental dos candidatos e ao aspecto pessoal. Para se criar uma tabela dinâmica no LibreOffice basta selecionar os dados, incluíndo os cabeçalos, e em seguida selecionar “dados” e “tabela dinâmica”. Em seguida seleciona-se como organizar os dados e seus dados centrais. Estas tabelas permitem pivoteamento, drill-down e drill-up.

5. Resultados

Com a ferramenta dOLAP por meio de tabelas dinâmicas foram criados uma série de pivôs que serviram de base para criação de gráficos objetivando responder as perguntas. Alguns destes gráficos seguem abaixos:

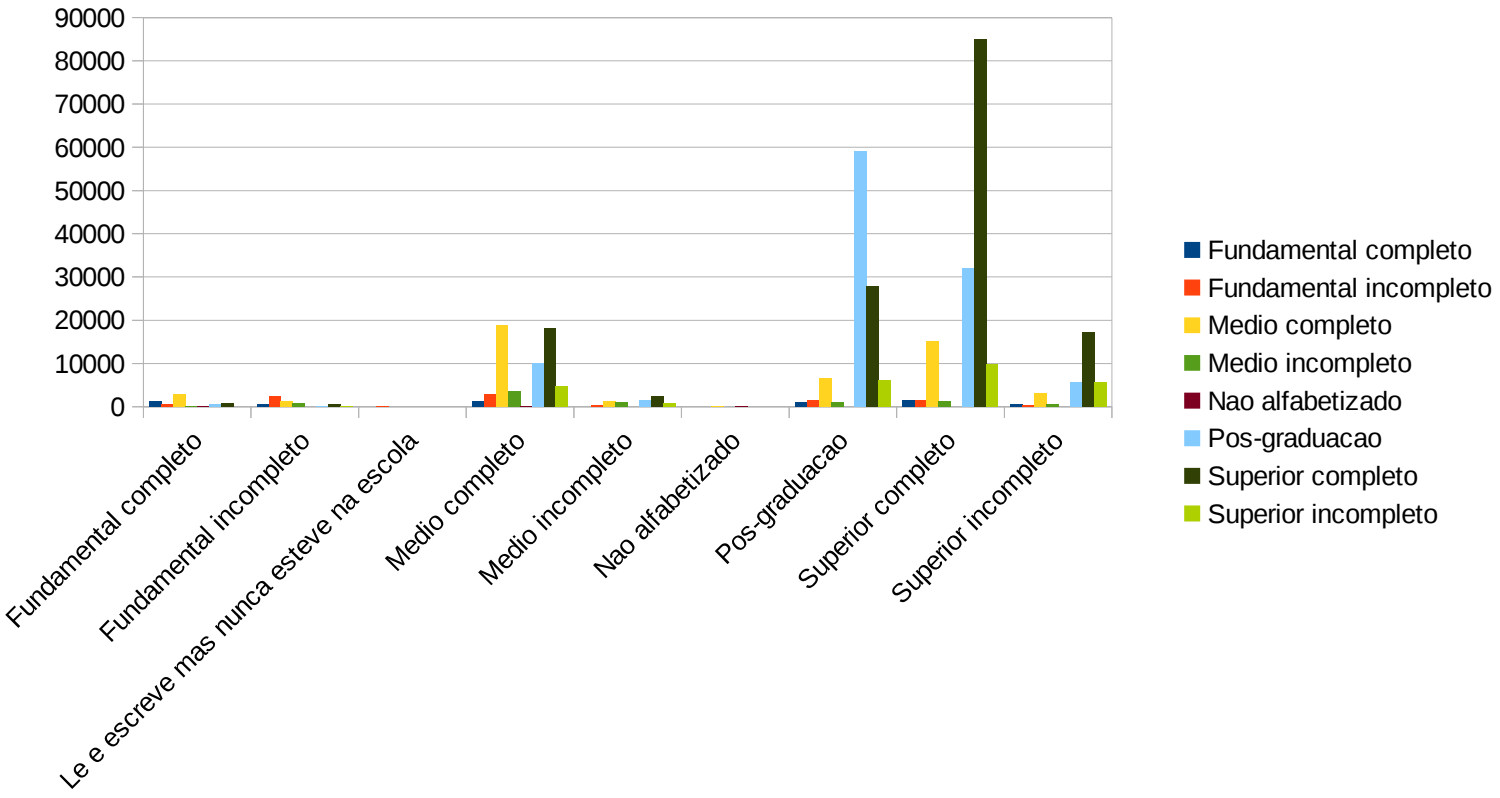


Gráfico 1: influencia da formação dos pais por total de pontuação

O gráfico 1 possuí a influência dos pais onde o eixo Y caracteriza a pontuação atingida e o eixo X possui a formação da mãe cruzada com a formação do pai. Apesar de a maior pontuação total aparecer quando pai e mãe possuem ensino superior completo, o Gráfico 2 deixa claro que a formação dos pais pouco influencia na média da pontuação, o que leva a crer que pais com ensino superior tendem a incentivar ou cobrar mais de seus filhos a formação.

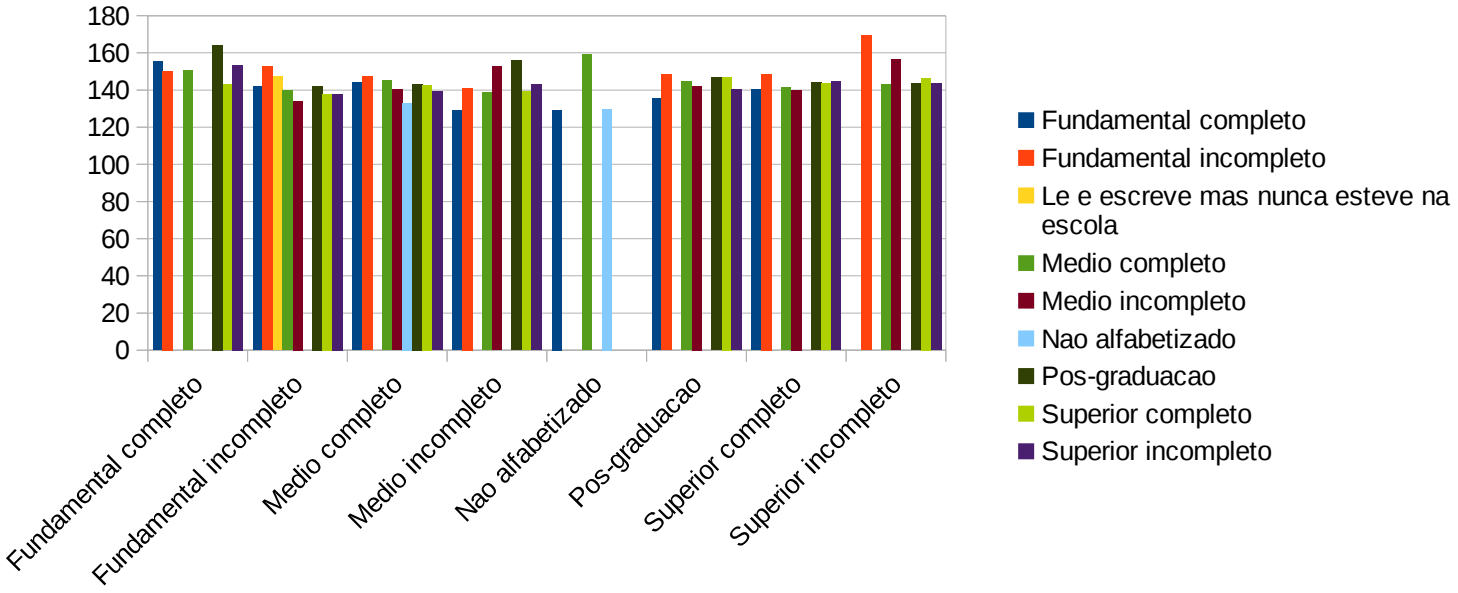


Gráfico 2: influencia da formação dos pais em média de pontuação.

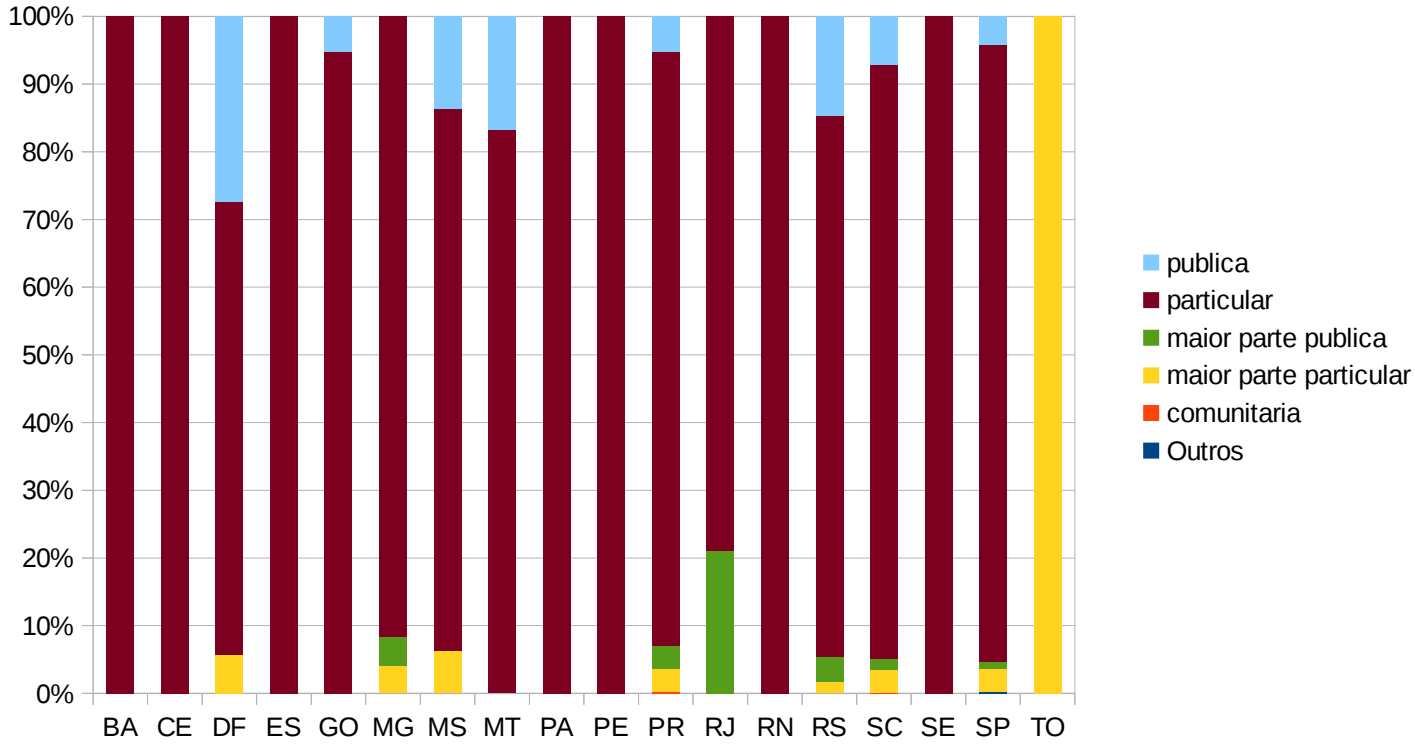


Gráfico 3: percentual de pontuação por ensino e por unidade da federação

No Gráfico 3 fica evidenciado que estudar em escolas particulares é fator característico e dominante , entretanto o Gráfico 4 demonstra que a média é mais influenciada pelo estado de origem do que pelo tipo de ensino, mesmo este tanto tendo alguma influencia.

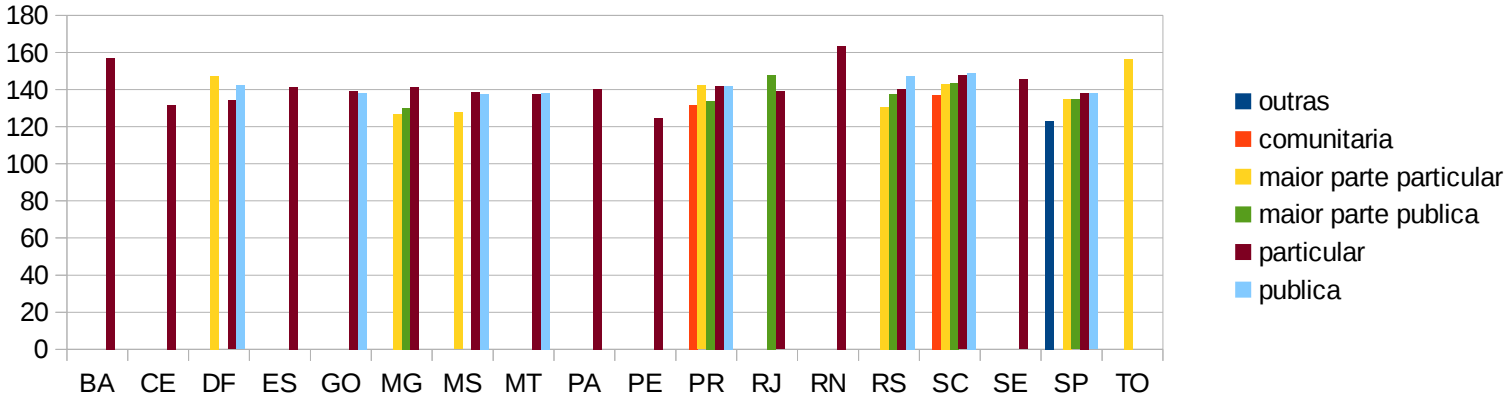


Gráfico 4: média de pontuação por ensino e por unidade da federação

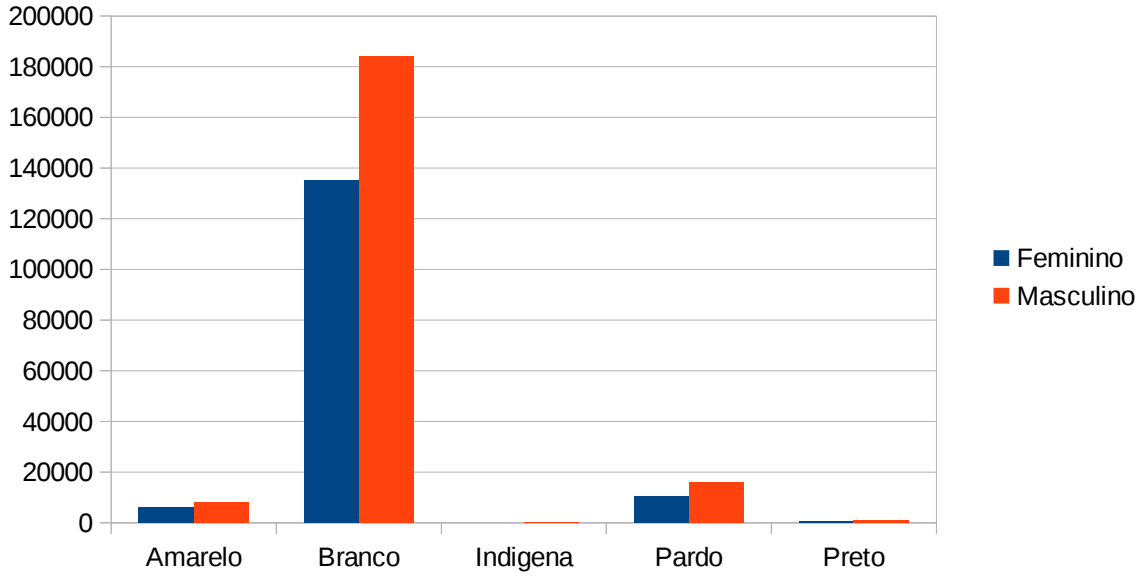


Gráfico 5: Pontuação total por raça e sexo.

Percebe-se que há uma quantidade maior de aprovados do sexo masculino e mais especificamente de pessoas Brancas por meio do Gráfico 5. O Gráfico 6 aponta que apesar de a dominância ser de homens brancos a média geral permanece quase inalterada.

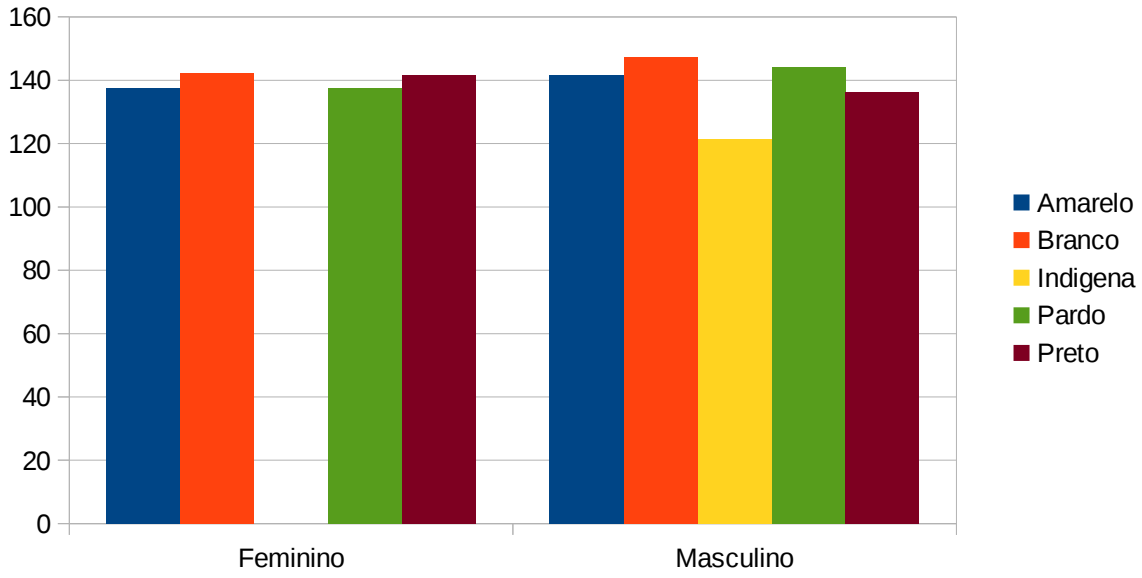


Gráfico 6: Percentual de pontuação por raça e ano.

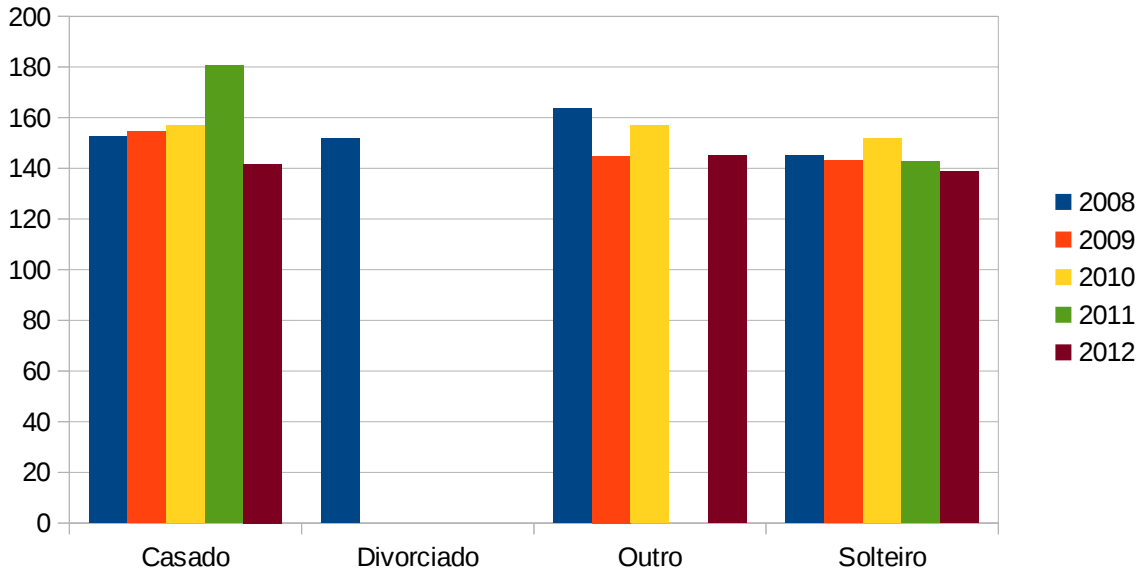


Gráfico 7: Situação civil dos candidatos por ano

Um dos objetivos do data mart desenvolvido é a capacidade de responder questões sobre as características dos candidatos que podem influenciar no sucesso. Para exemplificar uma possível exploração dos dados o Gráfico 7 ilustra a influencia do estado civil no desempenho do vestibular.

6. Conclusões e Trabalhos Futuros

Com o resultado obtido pode-se perceber que há uma dominância de características entre os candidatos com melhor aproveitamento mas observa-se também que a média geral é pouco influenciada. Apesar de haver pouca influencia o padrão para se atingir a melhor nota é homem branco, casado, proveniente de escola particular.

A ferramenta OLAP proposta entretanto carece de melhorias e um desenvolvimento melhor planejado e executado.

7. Referencias

Wiki OpenOffice,
https://wiki.openoffice.org/wiki/Database/Drivers/MySQL_Native/1.0 acesso em julho 2014
https://wiki.openoffice.org/wiki/Connect_MySQL_and_Base acesso em julho 2014

Sheepdog Guides from Sheepdog Software,
http://sheepdogguides.com/srv/s0MySqlDoInst.htm acesso em julho 2014

Wiki Pentaho
http://wiki.pentaho.com/display/EAI/03.+Hello+World+Example acesso em julho 2014
http://wiki.pentaho.com/display/EAI/01.+Installing+Kettle acesso em julho 2014
http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial acesso em julho 2014
http://wiki.pentaho.com/display/EAI/04.+Refining+Hello+World acesso em julho 2014

Adventures with Open Source BI,
http://type-exit.org/adventures-with-open-source-bi/2011/09/partitioning-in-kettle/ acesso em julho 2014