

# Projeto Final Data Warehouse

**Fernanda Hahn<sup>1</sup>, Henrique Saito<sup>2</sup>, Letícia Amarante<sup>3</sup>, Matheus Schaly<sup>4</sup>, Nicolas Antero<sup>5</sup>**

<sup>1</sup>Instituto de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)  
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brazil

**Resumo.** *Data warehouse é o recurso de apresentação consultável para os dados de uma empresa. Data mart é um subconjunto de um data warehouse e é normalmente orientado para uma linha de negócios ou equipe específica. O objetivo do trabalho consiste em analisar o modelo sócio acadêmico do vestibular da Coperve de 2008 a 2012 com vistas à implementação de um data mart para suporte e análises. As análises visam apoiar a secretaria do estado de SC na avaliação do desempenho nas disciplinas dos candidatos das escolas públicas do estado, comparando com as escolas privadas e federais. Para isso, elaboramos a modelagem dimensional (esquema estrela) e análises a serem feitas, realizamos o processo de ETL (extrair, transformar e carregar) utilizando, principalmente, os programas Spoon e HeidiSQL, e por último criamos o front-end (dashboard) utilizando a ferramenta PowerBI. O resultado obtido mostra que o desempenho dos alunos que frequentam escolas da rede privada e federal é melhor do que o desempenho dos alunos de escolas estaduais e municipais.*

## 1. Introdução

O problema consiste em fornecer suporte à tomada de decisão para a secretaria do estado de SC em relação a educação básica do estado. Para resolver o problema foi criado um sistema de apoio à decisão, neste caso, um data mart. Sistemas de apoio a decisão apoiam o processo de tomada de decisão e a definição de ações que auxiliam a compreensão dos resultados e a análise de tendências. A fim de auxiliar a tomada de decisão, criamos um dashboard que possibilita responder perguntas como: qual tipo de escola (privada ou pública) possuem as maiores notas no vestibular, qual tipo de escola é melhor/pior por matéria, cidades com maior/pior desempenho, entre outras. Além disso, os gráficos criados apresentam informações como: as redes de ensino (particular, municipal, estadual e federal) com as maiores médias por matéria, as piores e melhores escolas por matéria, a análise das escolas por microrregião e mesorregião, e permite a análise da evolução do desempenho anual das redes de ensino.

A justificativa para o problema descrito é a de aprimorar a tomada de decisão em relação a educação básica. Isso pode incluir, por exemplo, a melhor distribuição de recursos estatais para determinadas escolas públicas que estão obtendo um desempenho insatisfatório no vestibular. Melhorar a distribuição de recursos para cidades que estão com performance ruim. Criar políticas públicas que visem aprimorar o ensino em disciplinas que não estão atingindo as metas de desempenho. Criar ou aprimorar métricas de desempenho do ensino básico, como aumentar em 20% a nota na disciplina de matemática do vestibular até o ano 2025. Comparar a evolução do desempenho considerando diferentes disciplinas, cidades, tipos de escola, entre os anos de 2008 e 2012. Comparar o desempenho entre diferentes mesorregiões. Em suma, aprimorar o conhecimento acerca da educação do estado de SC.

## 2. Materiais

A Coperve (Comissão Permanente do Vestibular) controla os processos seletivos da UFSC, incluindo chamadas para os processos seletivos, programa de ações afirmativas (cotas), realização das provas, as diferentes formas de ingresso na UFSC, transferências e retornos de estudantes, entre outros assuntos. No presente artigo, estamos tratando do processo seletivo da UFSC através da realização do vestibular da UFSC.

O vestibular da UFSC é a principal porta de entrada aos cursos de graduação da instituição, que oferece cerca de 4.500 vagas em 101 cursos universitários. O Vestibular da UFSC é composto por três provas que são realizadas em três dias diferentes. Na primeira avaliação, o candidato deve realizar provas de primeira língua (12 questões), segunda língua (8 questões), matemática (10 questões) e biologia (10 questões). Na segunda prova, os seguintes temas são cobrados: ciências humanas e sociais (20 questões), sendo: história (7 questões), geografia (7 questões), filosofia (2 questões), sociologia (2 questões), questões interdisciplinares (2 questões), além disso há física (10 questões) e química (10 questões). Na terceira e última avaliação, o candidato deve fazer uma prova de redação composta por quatro questões discursivas (1).

O banco de dados relacional no qual o processo de ETL foi realizado possui 20 tabelas: cidade, unidade federativa, sexo, raça, boletim de desempenho, estabelecimento de ensino, candidato, candidato classificado, ponto candidato, grade socioeconômica, evento, local, língua estrangeira, acertos questões curso, disciplina prova, código questionário, área do curso, centro do curso e curso. Para o data mart proposto, modelado como uma esquema estrela, apenas 4 tabelas foram criadas, sendo elas 1 tabela de fato e 3 tabelas dimensões (evento, tempo e estabelecimento de ensino). Além do arquivo dump (.sql) do banco de dados, também temos uma tabela .xlsx contendo informações de 159.556 linhas (candidatos) e 19 colunas, uma tabela .xls de 1576 linhas e 4 colunas contendo os códigos dos questionários, uma tabela .xls de 171 linhas e 5 colunas contendo a grade socioeconômica, e um arquivo .erl contendo o modelo entidade relacionamento.

## 2.1. Métodos

O data warehouse é um tipo de sistema de apoio a decisão, é a fonte de dados consultável na empresa. O data warehouse nada mais é do que a união de todos os data marts constituintes. Um data warehouse é alimentado a partir da área de preparação de dados. O gerente do data warehouse é responsável tanto pelo data warehouse quanto pela área de preparação de dados (processo de ETL) (2). ETL é um tipo de processo de integração de dados que se refere a três etapas distintas, mas inter-relacionadas (extrair, transformar e carregar) e é usado para sintetizar dados de várias fontes muitas vezes para construir um data warehouse, data hub ou data lake (3).

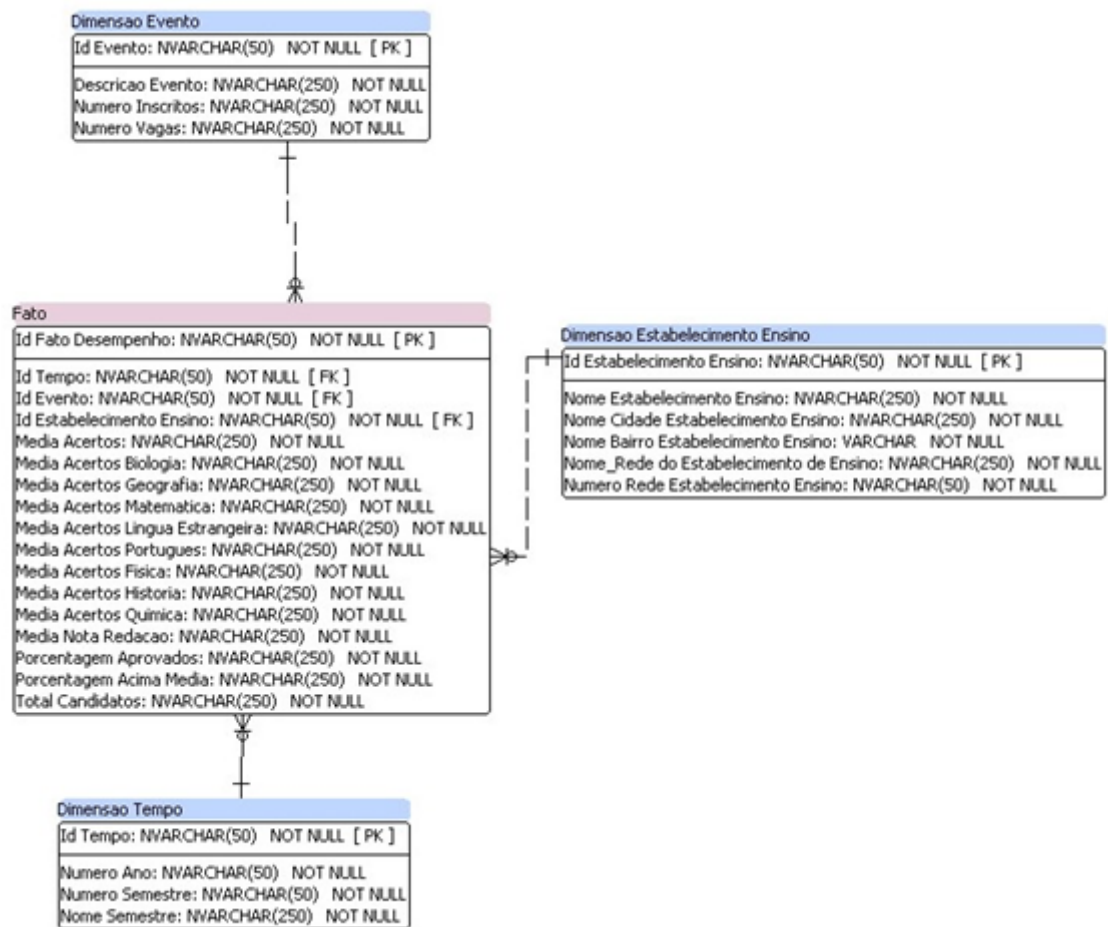
Especificamente, o data warehouse é o recurso de apresentação consultável para os dados de uma empresa e este recurso de apresentação não deve ser organizado em torno de um modelo de relação de entidade porque, ao usar a modelagem de relação de entidade, perderá a compreensibilidade e o desempenho. Além disso, o data warehouse é frequentemente atualizado em uma base de carga controlada, conforme os dados são corrigidos, os instantâneos são acumulados e os status e rótulos são alterados. Finalmente, o data warehouse é precisamente a união de seus data marts constituintes (2).

A modelagem dimensional é uma técnica de design lógico que busca apresentar os dados em uma estrutura padrão que é intuitiva e permite acesso de alto desempenho. É inerentemente dimensional e segue uma disciplina que usa o modelo relacional com algumas restrições importantes. Cada modelo dimensional é composto de uma tabela com uma chave multipartes, chamada tabela de fatos, e um conjunto de tabelas menores chamadas tabelas de dimensões. Cada tabela de dimensão possui uma chave primária de parte única que corresponde exatamente a um dos componentes da chave multiparte na tabela de fatos. Essa estrutura característica em forma de estrela costuma ser chamada de esquema estrela (star-schema) (2).

Data mart é um subconjunto lógico do data warehouse completo. Um data mart representa um projeto que pode ser concluído em vez de ser um empreendimento impossível. Um data warehouse é formado pela união de todos os seus data marts. Além dessa definição lógica bastante simples, geralmente vemos o data mart como a restrição do data warehouse a um único processo de negócios ou a um grupo de processos de negócios relacionados voltados para um determinado grupo de negócios. O data mart é provavelmente bancado e construído por uma única parte do negócio, e um data mart é geralmente organizado em torno de um único processo de negócio (2).

No presente trabalho construímos um sistema de apoio de decisão para um processo fim específico: apoiar a secretaria do estado de SC na tomada de decisão em relação a educação do estado. Sendo assim, a partir de um banco de dados relacional, foi elaborado a modelagem dimensional (esquema estrela), o processo de ETL, a criação de um data mart e um front-end (dashboard) para a utilização do sistema.

### 3. Metodologia



**Figura 1. Modelagem dimensional**

Para o processo de ETL os programas utilizados foram Pentaho Data Integration (Spoon), UniController, SQL Power Architect, Engenharia Reversa, HeidiSQL. A suite Pentaho é formada por um conjunto de softwares voltados para construção de soluções de BI (business intelligence) de ponta-a-ponta, que inclui programas para extrair os dados de sistemas de origem em uma empresa, gravá-los em um data warehouse (ou base de dados), limpá-los, prepará-los e entregá-los a outros sistemas de destino ou mesmo a outros componentes da suite para estudar ou dar acesso aos dados ao usuário final.

BI

information\_schema10,0 KiB

dwufsc240,0 KiB

dim\_candidato16,0 KiB

dim\_escola160,0 KiB

dim\_vestibular16,0 KiB

fato\_desempenho48,0 KiB

mysql

performance\_schema

phpmyadmin

vestibular

dwufsc.dim\_candidato: 134.161 registros totais (aproximadamente), limitado em 1.000

Próximo

Mostrar todos

Ordem

Colunas (7/7)

Filtro

Candidato_key	Id_Candidato	Tp_Lingua_Estrangeira	Ano_Segundo_Grau	Sexo_Candidato	Raca_Candidato	Cidade_Candidato
1	2000016	4	1999	F	1	JARAGUA DO SUL
2	2000024	4	2004	F	1	FLORIANOPOLIS
3	2000059	4	2002	M	1	FLORIANOPOLIS
4	2000067	4	2003	F	1	SUMARE
5	2000075	4	2003	M	1	JACINTO MACHADO
6	2000083	2	2002	M	1	FLORIANOPOLIS
7	2000091	2	2003	F	1	PALHOCA
8	2000105	4	2007	F	1	VIDEIRA
9	2000113	4	1998	M	1	FLORIANOPOLIS
10	2000121	4	1989	F	1	FLORIANOPOLIS
11	2000130	2	2006	F	1	SAO PAULO
12	2000164	4	1992	M	1	FLORIANOPOLIS
13	2000172	2	2007	F	1	CRISSIUMAL
14	2000180	2	2007	M	1	PALHOCA
15	2000199	2	2006	F	1	FLORIANOPOLIS
16	2000202	4	1996	F	1	BALNEARIO CAMBORIU

**Figura 2. Dados candidato**

O Pentaho Data Integration é o componente da suíte Pentaho usado para criar processos de ETL que alimentam o banco de dados. O Pentaho Data Integration é formado por duas categorias de artefatos, Jobs e Transformações, e estes artefatos são construídos por meio de sua interface gráfica, o Spoon. O Spoon é a interface gráfica do Pentaho Data Integration que facilita na concepção de rotinas e lógica ETL (4).

BI

information\_schema

10,0 KiB

dwufsc

240,0 KiB

dim\_candidato

16,0 KiB

dim\_escola

160,0 KiB

dim\_vestibular

16,0 KiB

fato\_desempenho

48,0 KiB

mysql

performance\_schema

phpmyadmin

vestibular

dwufsc.dim\_escola: 1.099 registros totais (aproximadamente), limitado em 1.000

Próximo

Mostrar todos

Ordem

Colunas (8/8)

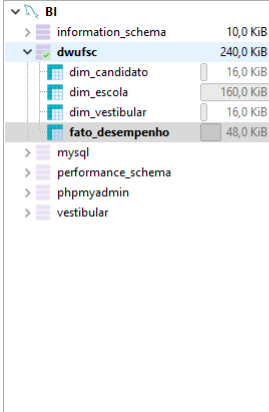
Filtro

Escola_key	Id_Escola	Nome_Escola	Nome_Cidade_Escola	Numero_Rede_Escola	Nome_Rede_Escola	Microrregiao_Escola
1	1	EEB JOSÉ ZANCHETTI	ABDON BATISTA	2	Rede Estadual	Curitibanos
2	10	EEB IRINEU BORNHAUSEN	AGUAS DE CHAPECO	2	Rede Estadual	Chapecó
3	100	EEB DOM JOAQUIM	BRACO DO NORTE	2	Rede Estadual	Tubarão
4	1000	COLÉGIO CENECISTA DR JÚLIO CÉSAR R NEVES	CONCORDIA	4	Rede Privada	Concórdia
5	1001	SENAI CENTRO DE ED. E TECN. DE CONCÓRDIA	CONCORDIA	4	Rede Privada	Concórdia
6	1003	CURSO E COLÉGIO GENIUS	FLORIANOPOLIS	4	Rede Privada	Florianópolis
7	1004	COLÉGIO UNIFICADO	GASPAR	4	Rede Privada	Blumenau
8	1005	EEB PROFª SALETE SCOTTI DOS SANTOS	ICARA	2	Rede Estadual	Criciúma
9	1006	SENAI-CENTRO DE TECN ELETROMETALMECÂNICA	JOINVILLE	4	Rede Privada	Joinville
10	1008	COLÉGIO SINERGIA	NAVEGANTES	4	Rede Privada	Itajaí
11	1009	EEB PROF BENONIVIO JOÃO MARTINS	PALHOCA	2	Rede Estadual	Florianópolis
12	1010	CENTRO EDUCACIONAL SÃO JUDAS TADEU	PALHOCA	4	Rede Privada	Florianópolis
13	1011	COLÉGIO ENERGIA	PALHOCA	4	Rede Privada	Florianópolis
14	1013	SENAI-CENTRO DE EDUCAÇÃO E TECNOLOGIA	RIO DO SUL	4	Rede Privada	Rio do Sul
15	1014	COLÉGIO DÓ-RÉ-MI	SANTO AMARO DA IMPERATRIZ	4	Rede Privada	Florianópolis
16	1015	EEB OSCAR MAJOLO	SAO MIGUEL DA BOA VISTA	2	Rede Estadual	Chapecó
17	1016	EXATHUM CURSO E COLÉGIO	JOINVILLE	4	Rede Privada	Joinville
18	1017	CENTRO EDUCACIONAL ATLÂNTICO SUL	BALNEARIO CAMBORIU	4	Rede Privada	Itajaí
19	1018	SENAI CTV BLUMENAU	BLUMENAU	4	Rede Privada	Blumenau
20	1019	EEM YVONE OLINGER APPEL	BRUSQUE	2	Rede Estadual	Blumenau

**Figura 3. Dados escola**

HeidiSQL permite ver e editar dados e estruturas de computadores rodando um dos sistemas de banco de dados MariaDB, MySQL, Microsoft SQL, PostgreSQL e SQLite.

Para o ETL, primeiramente foi realizado uma conexão MySQL local usando o UniController. Já com a conexão criada, os dados do dump sql, que contém o modelo entidade relacionamento, foram lidos utilizando o programa HeidiSQL. No HeidiSQL foi possível acessar essa conexão e criar um banco de dados chamado dwufsc. Em seguida a modelagem anteriormente desenvolvida no draw.io foi refeita no SQL Power Architect. Utilizando a ferramenta Engenharia Reversa, o modelo desenvolvido no SQL Power Architect foi utilizado para gerar o código SQL. O código SQL foi então usado como script no HeidiSQL para criar as tabelas de dimensões e a tabela de fato no banco de dados dwufsc anteriormente criado.



dwufsc.fato\_desempenho: 34.560 registros totais (aproximadamente), limitado em 1.000

didato_key	Vestibular_key	Escola_key	Acertos_Biolo...	Acertos_Geografia	Acertos_Matemati...	Acertos_Lingua_Estr...	Acertos_Por...	Acertos_Fi...	Acertos_Hi...	Acertos_Quim...
22.750	1	1	2	3	2	4	5	1	3	3
23.784	1	1	2	7	4	3	4	4	7	4
24.422	1	1	3	7	1	2	6	6	6	3
24.880	1	1	3	5	2	3	5	0	4	2
30.577	2	1	4	4	2	3	3	1	6	2
31.132	2	1	5	6	4	8	3	8	6	7
23.607	1	2	0	4	3	6	6	2	2	3
10.758	1	3	3	8	3	6	5	4	6	5
12.093	1	3	2	6	1	3	8	3	5	0
12.572	1	3	7	8	4	6	8	5	6	5
13.217	1	3	1	6	4	6	7	5	5	4
14.625	1	3	2	5	3	6	4	3	3	3
15.591	1	3	2	4	1	1	4	4	5	1
17.067	1	3	3	7	4	4	4	5	4	2
17.461	1	3	2	8	6	8	8	4	8	4
18.598	1	3	1	2	3	2	1	2	2	1
21.150	1	3	1	1	1	1	1	1	2	1
26.503	1	3	0	1	0	1	1	2	3	0
26.522	1	3	4	5	5	6	4	4	3	3

**Figura 4. Dados fato**

Já com a modelagem e banco de dados criados no HeidiSQL, foi utilizado o Spoon (do Pentaho Data Integration) para conectar na mesma conexão SQL do dwufsc e dos dados do vestibular. Além disso, foi necessário realizar alterações no script properties do Pentaho a fim de fazer as conexões funcionarem. No Spoon foi usado a conexão para puxar os dados do vestibular, executar as transformações necessárias e em seguida fazer a carga nas respectivas dimensões do banco de dados dwufsc.



dwufsc.dim\_vestibular: 5 registros totais (aproximadamente)

Vestibular_key	Desc_Vestibular	Id_Vestibular	Ano_Vestibular	Numero_Inscritos	Numero_Vagas
1	Vestibular 2008	15	2007	30612	4095
2	Vestibular 2009	16	2008	30854	4581
3	Vestibular 2010	20	2009	32524	6021
4	Vestibular 2011	25	2010	34876	5881
5	Vestibular 2012	28	2011	30358	5991

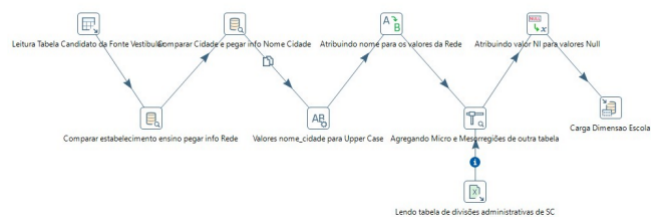
**Figura 5. Dados vestibular**

Ao realizar o carregamento de dados para tabela fato, ocorreu um problema devido à falta de memória RAM do computador e foi possível carregar apenas uma fração dos dados. Após diversas tentativas, concluímos que não seria possível carregar os dados naquela máquina e optamos por trabalhar apenas com o que tínhamos disponível.

Para o front-end, dashboard, foi utilizado a ferramenta PowerBI. O Power BI é uma coleção de serviços de software, aplicativos e conectores que funcionam juntos para transformar as fontes de dados não relacionadas em percepções coerentes, visualmente imersivas e interativas. Seus dados podem ser uma planilha do Excel ou uma coleção de data warehouses híbridos baseados em nuvem e no local. O Power BI permite facilmente conectar diferentes fontes de dados, visualizar e descobrir o que é importante nos seus dados (5).



**Figura 6. carga dimensao candidato**



**Figura 7. carga dimensão escola**



**Figura 8. carga dimensão vestibular**



**Figura 9. Carga dimensão fato**

## 4. Power BI

Para a representação das medidas foi escolhido o PowerBI. Primeiramente a equipe conectou o Data Mart ao Power BI e, após, utilizou a própria ferramenta para preparar a visualização dos dados. Foram criados, então, 5 gráficos interativos que utilizaram a média acumulada das diferentes redes de ensino.

Apresentam-se os gráficos referente às perguntas propostas.

1. Apresentar as redes de ensino (particular, municipal, estadual e federal) com as maiores médias por matéria. Na figura abaixo, observamos que a rede privada e a rede federal possuem maior desempenho, em todas as matérias, do que as redes estaduais e municipais. Também notamos que as maiores médias são das matérias de geografia, história, literatura e português, em todas as redes de ensino.

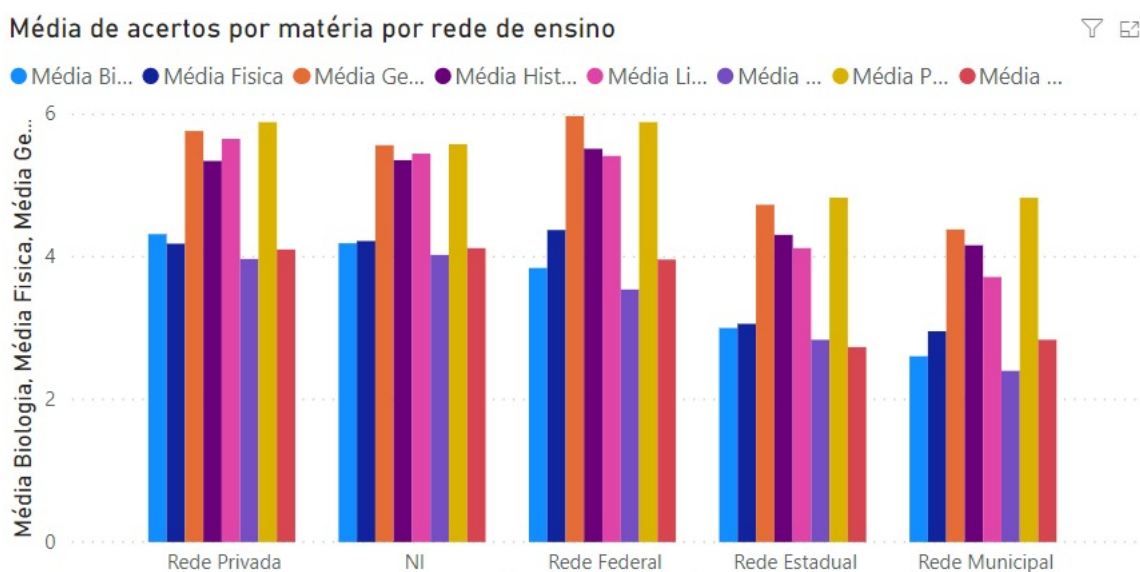


Figura 10. Gráfico referente as maiores médias por matéria das redes de ensino

2. Realizar a contagem das escolas por microrregião. Na figura abaixo, temos que a contagem de escolas por microrregião. Notamos que a cidade de Florianópolis representa a cidade com maior número de escolas, 151 escolas, representando 13.74% do total de escolas do estado.



Contagem de Nome\_Escola por Microrregiao\_Escola

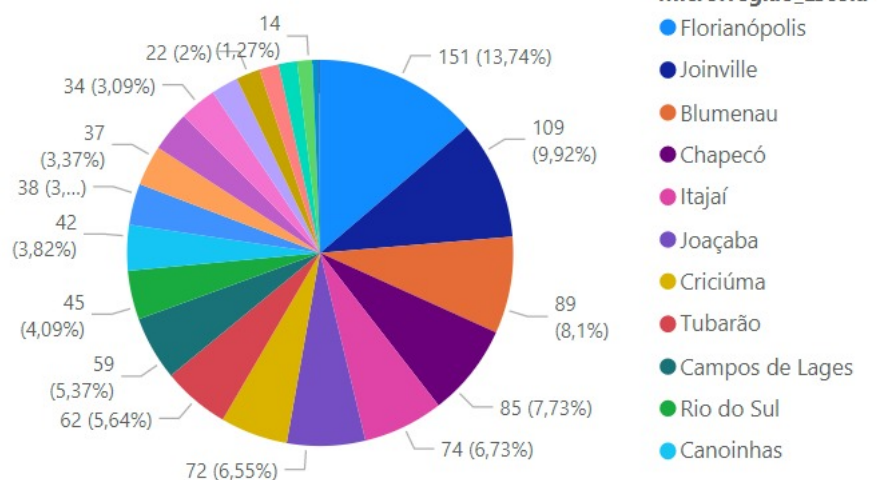


Figura 11. Gráfico referente as piores e melhores escolas por matéria

3. Definir as piores e melhores escolas por matéria. Na figura abaixo temos parte da uma tabela que contém o nome da escola seguido pela média em cada disciplina que os alunos de tal escola obtiveram no vestibular da UFSC.

Média Biologia	Média Historia	Média Fisica	Média Lingua_Estrangeira	Média Matematica	Média Portugues	Média Quimica	Média Geografia	Nome_Escola
8,00	8,00	7,00	8,00	7,00	9,00	7,00	9,00	CENTRO EN
7,00	6,00	5,00	4,00	7,00	5,00	5,00	6,00	C CENECIST.
7,00	5,00	4,00	8,00	7,00	7,00	3,00	7,00	EEB GOV HI
7,00	7,00	6,00	8,00	5,00	9,00	3,00	6,00	EEB PROF B.
7,00	5,00	3,00	9,00	6,00	4,00	6,00	4,00	NÚCLEO MU AGOSTINI
6,25	7,00	7,00	7,00	7,00	8,00	6,50	8,25	COLÉGIO DI
6,04	6,74	6,00	6,35	6,00	6,96	5,52	6,70	COLÉGIO SÂ
6,00	7,50	5,00	7,50	5,50	7,50	5,50	8,00	C CENECIST.
6,00	7,50	4,50	4,50	4,00	8,00	5,50	6,00	C TÉCNICO.
6,00	4,00	6,00	8,00	5,00	7,00	5,00	7,00	COLÉGIO DI
6,00	9,00	3,00	8,00	1,00	6,00	6,00	5,00	EEB ADOLFO
6,00	6,00	3,00	6,00	3,00	4,00	2,00	4,00	EEB CÂNDIA
6,00	3,00	6,00	8,00	6,00	9,00	2,00	6,00	EEB MARTIN
6,00	5,00	3,00	2,00	2,00	2,00	5,00	1,00	EEB ROSINA
3,59	4,94	3,76	4,87	3,55	5,40	3,36	5,29	

Figura 12. Gráfico 1 referente desempenho das escolas por microrregião e mesorregião

Contagem de Nome\_Escola por Mesorregiao\_Escola

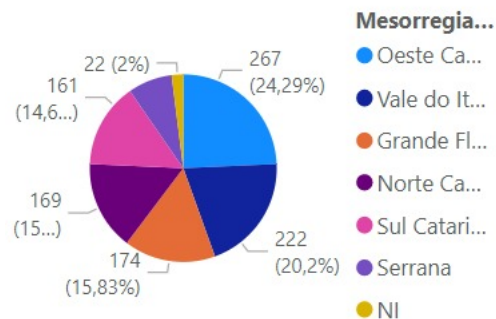


Figura 13. Gráfico 2 referente desempenho das escolas por microrregião e mesorregião

4. Realizar a contagem das escolas por mesorregião. Na figura abaixo, temos que a contagem de escolas por mesorregião. Notamos que a mesorregião do oeste representa a região com maior número de escolas, 267 escolas, representando 24.29% do total de escolas do estado. Enquanto que a região serrana possui apenas 22 escolas, representando 2% do número de escolas do estado.

Medida por Ano\_Vestibular e Nome\_Red\_Escola

Nome\_Red\_Escola ● NI ● Rede Estadual ● Rede Federal ● Rede Municipal ● Rede Privada

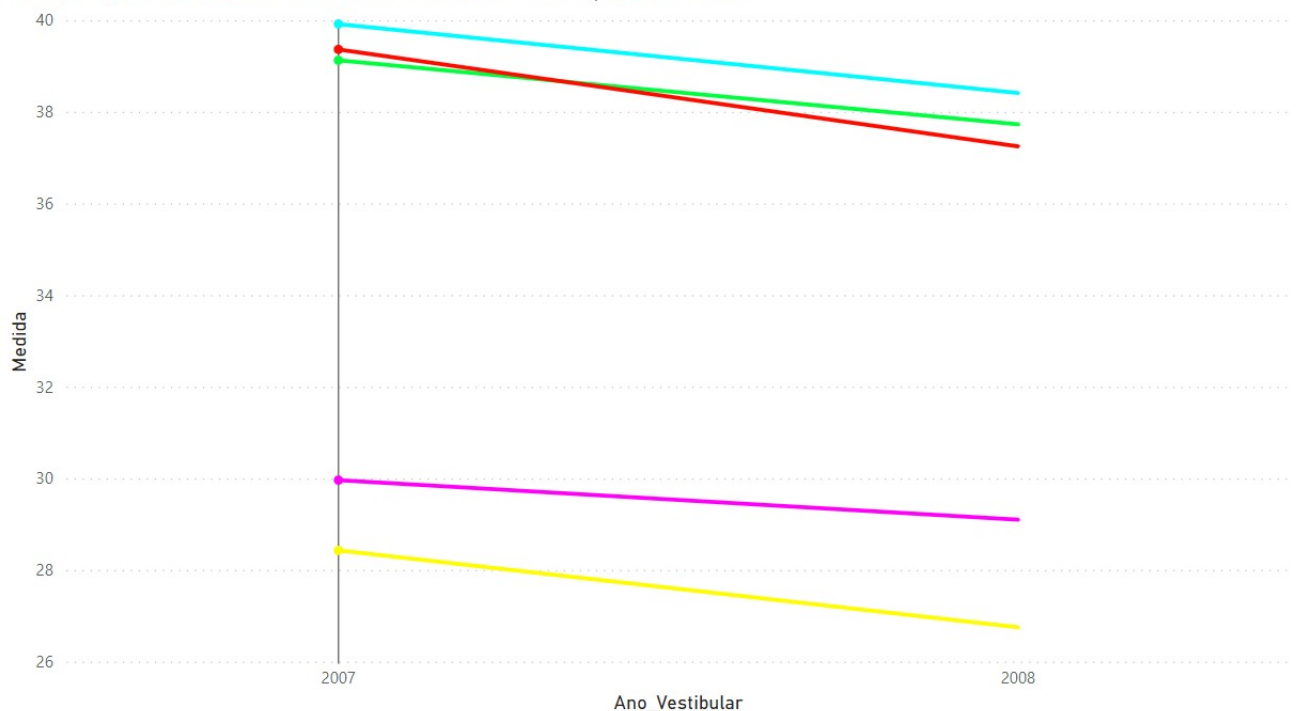


Figura 14. Gráfico referente a análise da evolução do desempenho anual das redes de ensino

## **5. Resultados**

A equipe conseguiu finalizar a implementação do Data Mart que foi construído de forma a contemplar os requisitos definidos. Apesar dos problemas com o carregamento dos dados, ainda foi possível utilizar 47895 linhas para a análise do cenário. Assim, após a construção do front-end com a ferramenta Power BI, obtivemos as respostas para os requisitos de forma visual. Assim, foi possível constatar que os gráficos representados eram bons retratos da realidade.

## **6. Conclusões e Trabalhos Futuros**

O trabalho analisou o modelo sócio acadêmico do vestibular da Coperve de 2008 a 2012 a fim de auxiliar a secretaria do estado de SC na avaliação do desempenho nas disciplinas dos candidatos das escolas públicas e privadas do estado. Um data mart foi criado para auxiliar na tomada de decisão. Os passos para a criação do data mart foram relatados, desde a modelagem dimensional, passando pelo ETL e resultando no front-end. Os resultados das análises obtidos também visam auxiliar o processo de decisão da secretaria do estado. Para trabalhos futuros, poder-se-ia relatar o treinamento fornecido ao usuário final sobre como utilizar o data mart. Além disso, poderia haver a observação contínua do data mart, demonstrando o processo contínuo de manutenção da base de dados. Também seria viável, como trabalho futuro, relatar o possível replanejamento do projeto e como se daria o processo de reestruturação de um data mart que já esteve anteriormente em produção.

## **7. Referências**

- (1) KIMBALL, Ralph. The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses.
- (2) <https://www.guiadacarreira.com.br/educacao/vestibular/vestibular-ufsc>.
- (3) <https://medium.com/hashmapinc/etl-understanding-it-and-effectively-using-it-f827a5b3e54d>.
- (4) <https://www.infoq.com/br/articles/pentaho-pdi/>.
- (5) <https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview>.