|  |  |
| --- | --- |
| marcaUerj_logo_coresPRETOeBRANCO | **Universidade do Estado do Rio de Janeiro**  Centro de Tecnologia e Ciências  Instituto de Matemática e Estatística |

Saulo Martins Soares da Fonseca

**Uso de DW para análise estatística do Enem na busca por melhoria educacional do país**

Orientadora: Prof.a Dra. Ana Carolina Brito de Almeida

Rio de Janeiro

2016

Sumário

[1.INTRODUÇÃO 3](#__RefHeading___Toc581_1837944507)

[2.MOTIVAÇÃO 3](#__RefHeading___Toc905_1837944507)

[3.OBJETIVO 3](#__RefHeading___Toc1765_1837944507)

[4.REVISÃO LITERÁRIA 4](#__RefHeading___Toc1132_519844130)

[5.METODOLOGIA 4](#__RefHeading___Toc1760_1837944507)

[6.RESULTADOS 6](#__RefHeading___Toc259_1626949292)

[7.CRONOGRAMA 7](#__RefHeading___Toc270_1626949292)

[8.REFERÊNCIAS 8](#__RefHeading___Toc326_1626949292)

[9.ASSINATURAS 9](#__RefHeading___Toc354_1626949292)

[Rio de Janeiro, X de dezembro de 2016 9](#__RefHeading___Toc358_1626949292)

# INTRODUÇÃO

Todo ano, em nível nacional, é realizado o Exame Nacional do Ensino Médio (ENEM) **¹** que tem por objetivo avaliar a qualidade do ensino médio e ser a porta de entrada para algumas universidades públicas e programas educacionais, tais como: Sistema de Seleção Unificada (SISU)², Programa Universidade para Todos (PROUNI)³, Financiamento Estudantil (FIES)4, Universidades em Portugal, Ciência sem Fronteiras e Certificação do Ensino Médio.

O ENEM abrange as matérias do currículo do ensino médio: Português (Literatura e Redação), Matemática, Geografia, História, Física, Química, Biologia e Língua Estrangeira. A prova é composta por uma redação e mais 180 questões objetivas distribuídas entre as áreas determinadas pelo Enem (Ciências da Natureza, Linguagens e Códigos, Ciências Humanas e Matemática). A prova deve ser realizada em dois dias em um prazo de até 5 horas.

Para o trabalho que será desenvolvido, na linha de pesquisa de Banco de Dados, o ENEM é uma fonte importante de dados, pois gera uma enorme quantidade de informações sobre a prova e o estudante que a realiza. A partir de tais dados, pode ser possível extrair informação útil para futuras análises sobre a educação nacional e os estudantes.

# MOTIVAÇÃO

A principal motivação é a oportunidade de aplicar os conhecimentos adquiridos durante o período da graduação, com enfoque na área de Banco de Dados.

Outra motivação é a oportunidade de tratar e disponibilizar as informações do Enem por um banco de dados para que os mesmos sejam usados de uma forma fácil para manipulação. Portanto, a principal motivação é transformar dados sem nenhum tipo de organização, sem nenhum tipo de tratamento, sem nenhum tipo de leitura possível e interpretação em dados organizados em tabelas organizadas, de fácil consulta e de fácil de interpretação.

# OBJETIVO

Os objetivos principais que a monografia contemplará e que foram definidos durante inúmeras reuniões são:

* Estender o MER (Modelo Entidade Relacional) e o modelo relacional da monografia do Erick [] para contemplar novas informações que podem ser relevantes para futuras análises.
* Implementar o banco de dados usando o SGBD relacional Postgresql.
* Elaborar scripts de carga dos arquivos textos para o banco de dados.
* Realizar a carga dos dados do ENEM 2011, 2012 e 2013.
* Identificar nova base de dados abertos que seja relevante para o cruzamento com os dados do ENEM
* Levantar consultas relevantes para o domínio educacional com os usuários.
* Gerar o modelo multidimensional, utilizando o esquema estrela [].
* Realizar o processo de Data Warehousing, implementando o modelo multidimensional, realizando a mineração de dados (Data Mining) entre outras etapas.
* Implementar as consultas que foram levantadas.
* Estudar ferramenta de visualização dos dados.
* Gerar gráficos a partir dos dados.
* Dar interpretação aos dados.

# REVISÃO LITERÁRIA

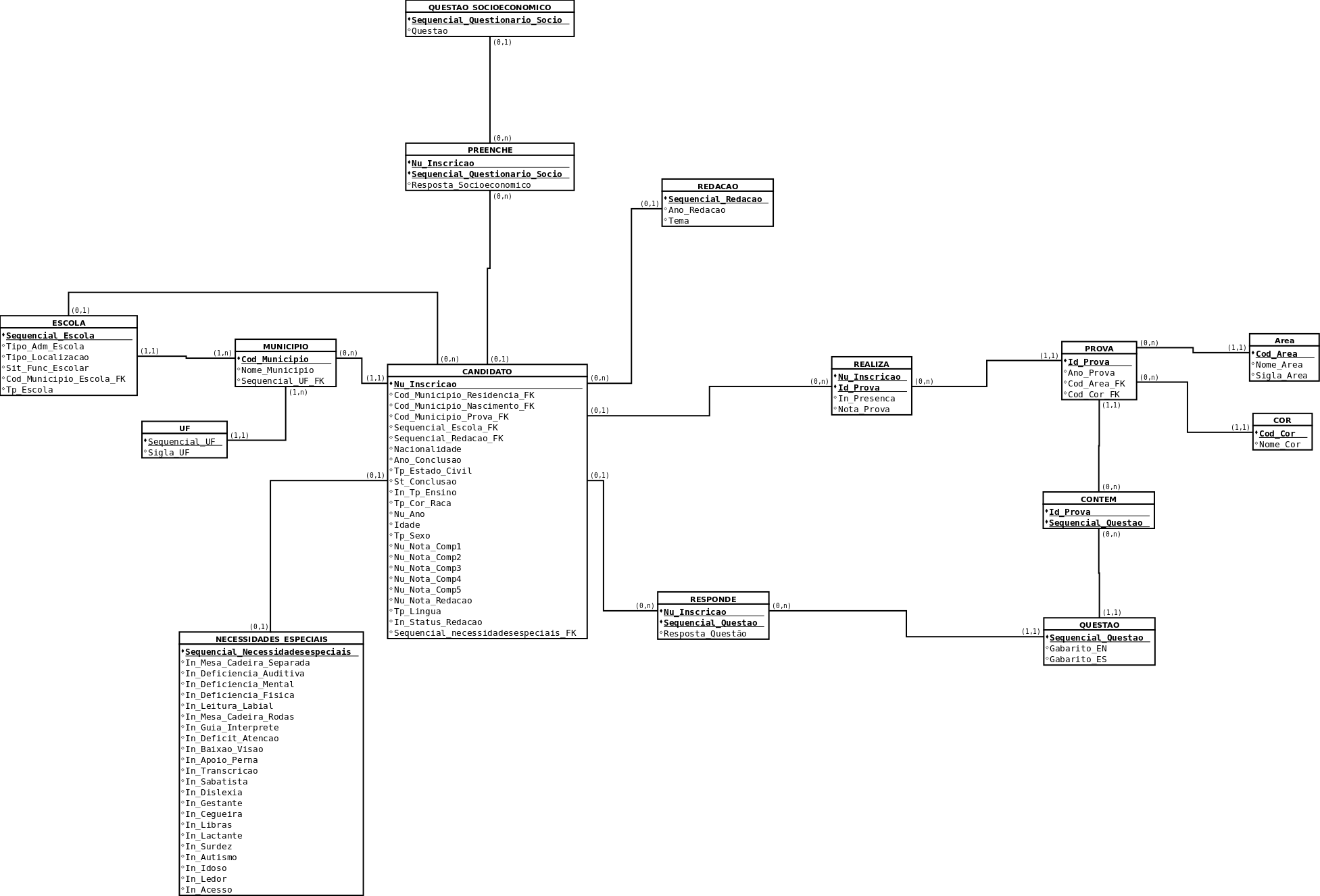
Para desenvolver a monografia, foram usados dois trabalhos como base: um trabalho de iniciação científica [Grinstein, 2015]5 e uma dissertação de mestrado (Frias, 2015)6, ambos alunos da PUC-Rio. Tais trabalhos foram escolhidos devido a uma parceria da UERJ com o grupo de pesquisa de sintonia fina de banco de dados da PUC-Rio.

Através dessa parceria, foram levantadas diversas ideiais. Entre elas, a extensão do MER para englobar as informações sobre o Questionário Socioeconômico, que não foi apreciado no escopo determinado no trabalho de iniciação científica do aluno Grinstein e pode ser uma fonte de análise futura. Além disso, no trabalho a ser desenvolvido, serão observados os dados dos anos de 2011 até 2014. Nos trabalhos levantados [][][] até o momento, só foi usado um único ano, impedindo a análise da evolução do ENEM ao longo dos anos. Por fim, os dados serão armazenados em banco de dados relacional, o que facilita a manipulação de dados e possibilita o armazenamento de grandes volumes de dados, diferente da ferramenta usada por [], que possui limitação de capacidade de linhas.

# METODOLOGIA

Para produzirmos o presente trabalho usaremos, principalmente, os dados do site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)7, onde os dados vem em arquivos nos formatos: “.txt” e “.csv”, porém os dados as vezes vem com alguns erros que são tratados com programas corretivos escritos em *Shell Script,* em Ansi C e Python.

Depois de todos os erros serem tratados os mesmos são enviados para o banco de dados Postgresql8 onde o esquema físico está pronto, respeitando fielmente o modelo ER feito no software DIA9 o qual podemos ver abaixo.

Figura 1 Modelo Relacional Proposto - ENEM

A Figura 1 reflete a organização e a interpretação dos dados do Enem na visão deste trabalho.

Em palavras resumidas, o modelo relacional é composto por quinze tabelas que o seu significado que será descrito abaixo:

* Entidade Escola: Compreende toda a informação da escola onde o inscrito estudou ou onde realizará o exame.
* Entidade Municipio: Compreende os códigos e nomes dos municípios.
* Entidade UF: Compreende as Unidades Federativas do Brasil.
* Entidade Candidato: Compreende, basicamente, todas as informações do candidato inscrito como: lugar onde nasceu, onde mora, notas das competências entre outras informações.
* Entidade Necessidades Especiais: Compreende todas as informações dos inscritos que são portadores de alguma necessidade especial ou não.
* Entidade Preenche: Compreende as respostas do Questionário Socioeconômico que o inscrito preencheu.
* Entidade Questao Socioeconomico: Compreende o enunciado das questões socioeconômicas.
* Entidade Redacao: Compreende o tema e ano da redação.
* Entidade Responde: Compreende as respostas dos quatro exames que o candidato realiza.
* Entidade Realiza: Compreende a presença do aluno e a nota dos quatro exames que o candidato realiza.
* Entidade Questao: Compreende o gabarito das provas.
* Entidade Contem: Compreende, para fins relacionais, as sequenciais das entidades: Prova e Questao.
* Entidade Prova: Compreende o ano da prova e os códigos das áreas e das cores dos exames.
* Entidade Area: Compreende o nome das áreas das provas e as siglas das mesmas.
* Entidade Cor: Compreende o nome da cor do exame,.

Outro ponto importante é que será gerado uma modelagem multidimensional que servirá de referencial para gerar os *Data Mart’s,* e assim na hora da mineração de dados eles sejam destinados aos escopos corretamente, por exemplo, um escopo só de escolas onde os dados tanto do Enem quanto os dados externos fiquem nesse ambiente e gerem estatísticas a serem aproveitadas.

# RESULTADOS

Os resultados esperados são os resultados das possibilidades de cruzamento de informações da base do ENEM com dados externos, ou seja, dados que não pertencem ao domínio do nosso trabalho, tais como: dados de investimentos em educação, dados de quantidade de escolas, quantidade de estudantes, quantidade de professores etc. A partir desse cruzamento, espera-se que surja informações estatísticas que possam ser analisadas e mostrem tendências sobre diversos assuntos na área da educação e possa trazer melhorias para as escolas e como consequencia, para a educação nacional.

Até o momento, só foram carregados os dados do ano de 2011 e um exemplo de consulta poderia ser para saber a quantidade de estudantes por estado do Brasil. A partir dos dados retornados, gerou-se um gráfico (vide Figura 2).

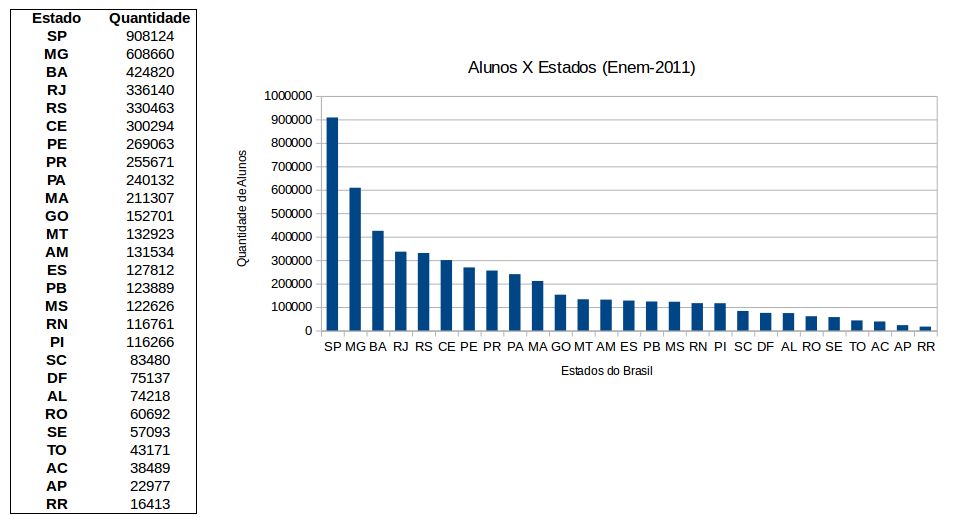


Figura 2 – Tabela e gráfico gerados a partir dos dados que foram gerados da base de dados.

Após isto, foi pesquisado o investimento em reais para educação no ano de 2011 e foi gerado a tabela abaixo (Figura 3).



Figura 3 – Tabela de investimento em educação e cultura (R$) por estado brasileiro no ano de 2011. Fonte dos dados: deepask10.

Com essas informações tanto internas quando externas, podemos observar a **tendência** de que os estados que tiveram mais investimentos no ano de 2011, foram os que mais tiveram inscritos no Enem.

# CRONOGRAMA

As atividades que foram elencadas até agora para realizar a monografia estão enumeradas abaixo junto com uma representação gráfica para facilitar a visualização, entretanto outras atividades podem ser acrescentadas ao longo do tempo.

1. Realizar pesquisa bibliográfica, buscando aprofundar o estado da arte em

Data Warehouse e trabalhos com o uso dos dados do ENEM;

2.

3. Estudar e analisar os dados do ENEM;

4. Estender o modelo de dados de [Grinstein, 2015] para contemplar novas

informações;

5. Implementar o modelo de dados em um banco de dados relacional;

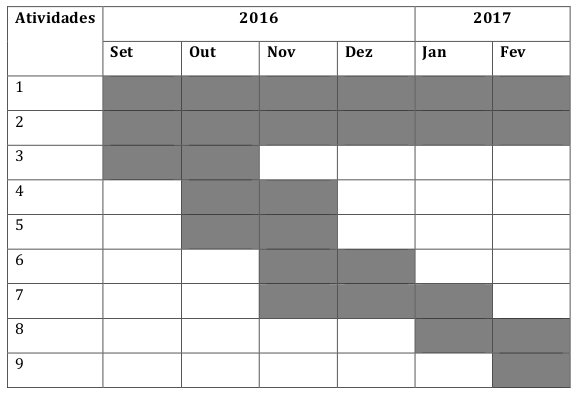
6. Elaborar scripts de carga do formato CSV do ENEM para o banco de dados

relacional (2013 e 2014);

7. Realizar a carga de dados no banco de dados relacional;

8. Elaborar o modelo de dados multidimensional;

9. Elaborar monografia.

Figura X – Modelo Entidade-Relacional refletindo a organização e a interpretação dos dados do Enem na visão deste trabalho.

# REFERÊNCIAS

1. Enem, disponível em: <http://enem.inep.gov.br/>, visitado em dezembro de 2016.

2. Sisu, disponível em: <http://sisu.mec.gov.br/>, visitado em dezembro de 2016.

3. Prouni, disponível em: <http://prouniportal.mec.gov.br/>, visitado em dezembro de 2016.

4. Fies, disponível em: <http://sisfiesportal.mec.gov.br/>, visitado em dezembro de 2016.

5. Grinstein, E. Z. **DBX: Ferramenta para apoio de sintonia fina de visões materializadas em bancos de dados relacionais/Um Banco de Dados para coleta de estatísticas do ENEM**, trabalho de iniciação científica apresentado ao departamento de informática da PUC-Rio, 2015.

6. Frias, J. L. D. **Uma ferramenta para a obtenção e análise de dados do ENEM**, dissertação de mestrado apresentada ao programa de pós-graduação em matemática da PUC-Rio, 2015.

7. Inep, disponível em: <http://portal.inep.gov.br/basica-levantamentos-acessar>, visitado em dezembro de 2016.

8. Postgresql, disponível em: <https://www.postgresql.org/>, visitado em dezembro de 2016.

9. DIA, disponível em: <http://dia-installer.de/>, visitado em dezembro de 2016.

10. DeepAsk, disponível em: <http://www.deepask.com.br/goes?page=Veja-ranking-estadual-pelo-investimento-em-educacao-e-cultura>, visitado em dezembro de 2016.

# ASSINATURAS

O pré-projeto foi desenvolvido pelo aluno de graduação Saulo Martins Soares da Fonseca sob a orientação da professora Prof.a Dra. Ana Carolina Brito de Almeida que seguirá orientando-o na monografia.

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Saulo Martins Soares da Fonseca | Ana Carolina Brito de Almeida |

Rio de Janeiro, X de dezembro de 2016