|  |  |
| --- | --- |
| 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, acesso a programas educacionais: 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 exame 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 que são: 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 em Banco de Dados o Enem é uma fonte importante, pois ela gera muitos dados informativos sobre a prova e o estudante que a realiza, por isso há o interesse em estudar essa fonte e se possível transformar em informação útil para ser realizadas 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, para isso foi escolhido o Enem como matéria de estudo para aplicar os conhecimentos que serão desenvolvidos ao longo da produção da monografia.

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 foram definidas durante inúmeras reunião são estas abaixo:

* Estender o MER (Modelo Entidade Relacional) e ER (Modelo Entidade-Relacional) da monografia do Erick.
* Construir o banco de dados usando o SGBD Postgresql.
* Realizar os inserts nas tabelas.
* Realizar as principais consultas.
* Gerar o modelo multidimensional.
* Gerar o Data Mart.
* Fazer o cruzamento de dados.
* Gerar gráficos a partir dos dados.
* Dar interpretação aos dados.

Todavia durante o período da elaboração da monografia os objetivos podem mudar ou serem acrescentados mais objetivos por causa das prioridades ao decorrer do tempo.

# REVISÃO LITERÁRIA

Para desenvolver a monografia dois trabalhos foram usados como base esses trabalhos foram: o primeiro trabalho é de iniciação científica (Grinstein, 2015)5 e o segundo é a dissertação de mestrado (Frias, 2015)6 ambos alunos da PUC.

Estes trabalham serviram de base para o desenvolvimento da monografia, portanto tanto deles foram tirados ideiais e acrescentados a futura monografia que será desenvolvida, por exemplo, da monografia foi estendido o MER para englobar o Questionário Socioeconômico que não foi apreciado no trabalho de iniciação científica do aluno Grinstein e serão observados os dados dos anos de 2011 até 2014 que só foi observado um único ano. Agora, em relação a dissertação será substituído a ferramentada do Excel para Banco de Dados, pois como os dados vem em várias linhas e o Excel possui limitação de capacidade de linhas, por isso essa ferramenta será substituída.

# METODOLOGIA

Para produzirmos a monografia 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.

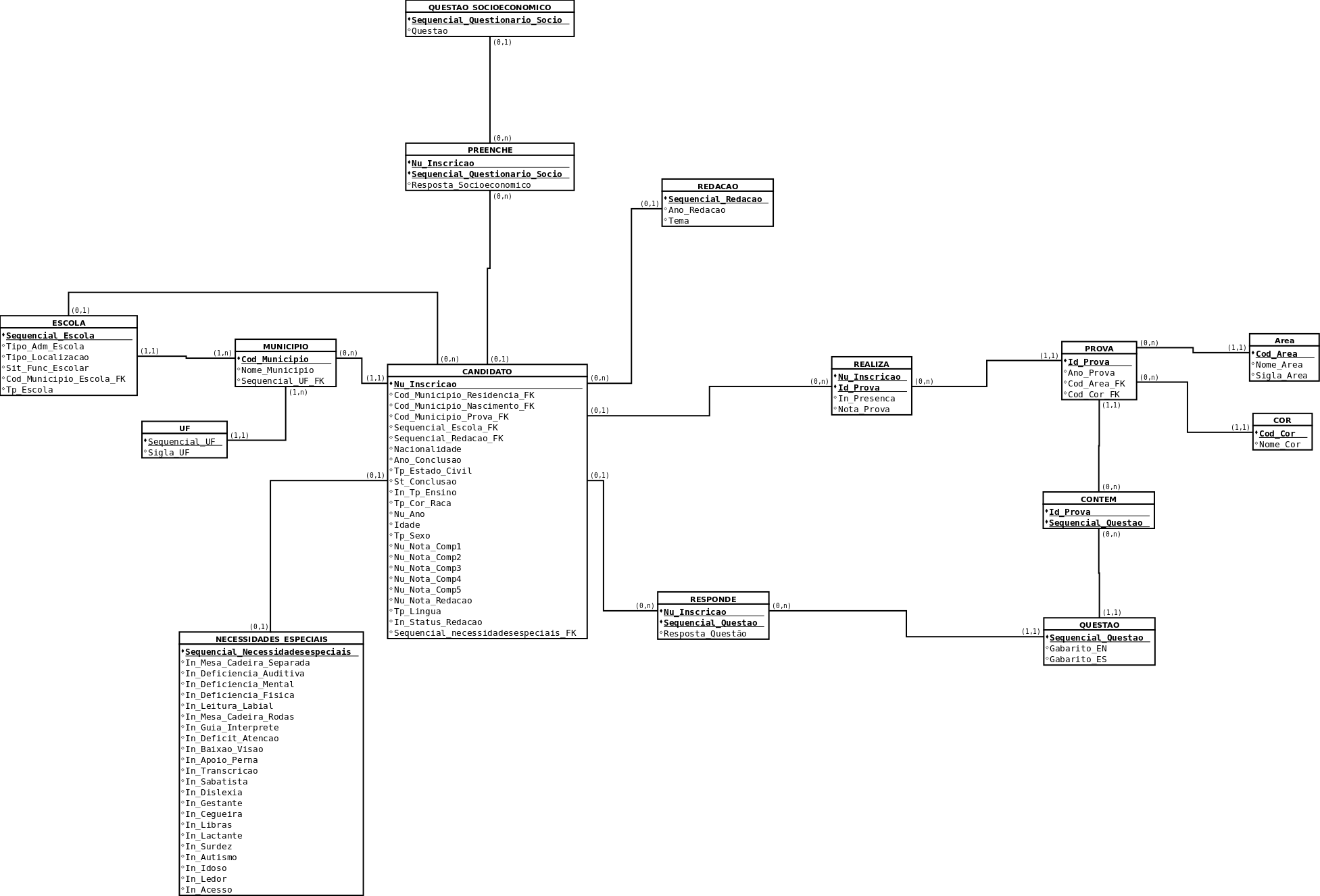


Figura 1 – Modelo Entidade-Relacional refletindo a organização e a interpretação dos dados do Enem na visão deste trabalho.

Em palavras resumidas, o modelo ER 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 do municípios, esta se relaciona com as entidades: Escola, Uf e Candidato.
* 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, esta entidade se relaciona com as entidades: Necessidades Especiais, Preenche, Redacao, Realiza e Responde.
* Entidade Necessidades Especiais: Compreende todas as informações dos inscritos que são portadores de alguma necessidade especial ou não, esta entidade se relaciona com a entidade Candidato.
* Entidade Preenche: Compreende as respostas do Questionário Socioeconômico que o inscritos preencheu, esta entidade se relaciona com as entidades: Candidato e Questao Socioeconomico.
* Entidade Questao Socioeconomico: Compreende o enunciado das questões socioeconômicas, esta se relaciona com a entidade Preenche.
* Entidade Redacao: Compreende o tema e ano da redação, esta entidade relaciona-se com a entidade Candidato.
* Entidade Responde: Compreende as respostas dos quatro exames que o candidato realiza, esta entidade relaciona-se com a entidade Candidato e Questao.
* Entidade Realiza: Compreende a presença do aluno e a nota dos quatro exames que o candidato realiza, esta entidade relaciona-se com com as entidades: candidato e prova.
* Entidade Questao: Compreende o gabarito das provas, esta entidade relaciona-se com com as entidades: Responde e Contem.
* Entidade Contem: Compreende para fins relacionais as sequenciais das entidades: Prova e Questao, esta entidade relaciona-se com as entidades ditas anteriormente.
* Entidade Prova: Compreende o ano da prova e os códigos das áreas e das cores dos exames, esta entidade relaciona-se com as entidades: Contem, Area e Cor.
* Entidade Area: Compreende o nome das áreas das provas e as siglas das mesmas, esta entidade relaciona-se com a entidade: Prova.
* Entidade Cor: Compreende o nome da cor do exame, esta entidade relaciona-se com a entidade: Prova.

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 com os dados da base com dados externos como, ou seja, dados que não pertencem ao domínio do nosso trabalho como: dados de investimentos em educação, dados de quantidade de escolas, quantidade de estudantes, quantidade de professores e etc, e desse cruzamento surjam informações estatísticas que possam ser analisadas e mostrem tendências sobre diversos assuntos na área da educação.

Por exemplo, foi realizado uma consulta na base de dados para saber a quantidade de estudantes por estado do Brasil e foi retornado uma tabela de dados que gerou um gráfico. Veja abaixo:

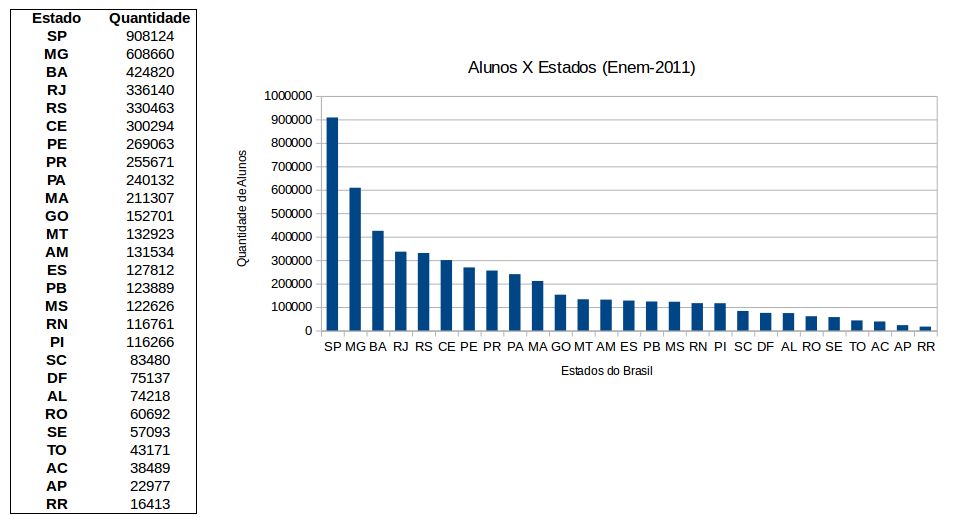


Figura 2 – Tabela e gráfico gerados a partir dos dados que foram gerados da base de dados.

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



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 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 acrescentados 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. Participar de discussões e execução de atividades em projetos de pesquisa

ou desenvolvimento tecnológico de forma a contribuir com as atividades e

crescimento da linha de pesquisa de Banco de Dados;

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

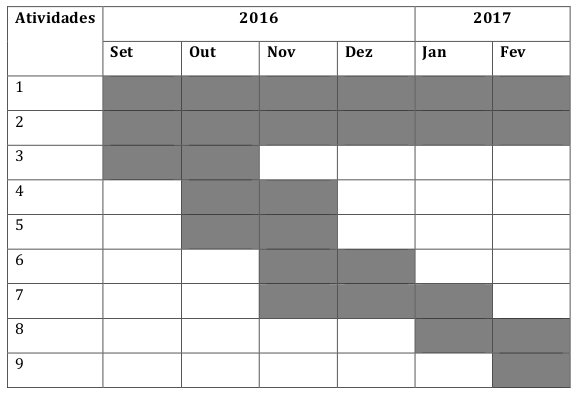
relacional (2013 e 2014);

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

8. Elaborar o modelo de dados multidimensional;

9. Produção e elaboração de relatório técnico, buscando ampliar o

conhecimento sobre este tema na UERJ.

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