**Documentação da Metodologia – Aplicação de Business Intelligence nos das dos discentes da Universidade Estadual do Norte do Paraná**

## 1 – Extração

Os dados foram extraídos por meio do Sistema Unificado de Administração pública(SUAP)

## 2 – Limpeza - Excel

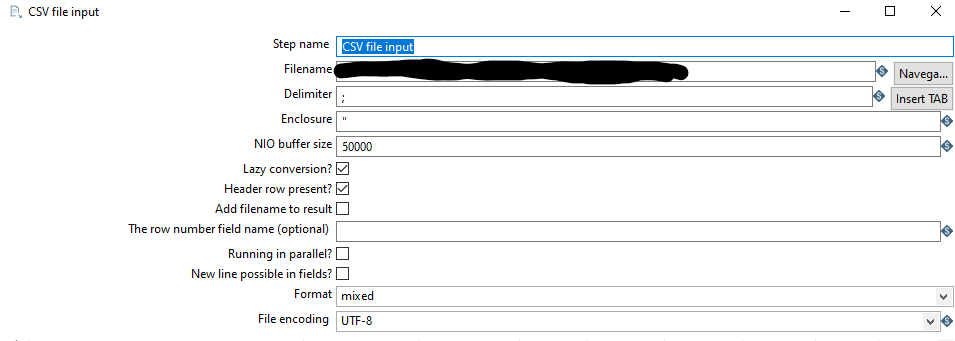
Uma vez extraídos, os dados foram sujeitos à um processo de limpeza na ferramenta *Microsoft Excel* num primeiro momento. Para se verificar se haviam Campos(Colunas) vazios, foi se utilizado o recurso “Filtro” na aba “Dados” da ferramenta, para se verificar quais os valores que preenchiam cada campo, caso os campos apresentassem somente valores nulos, vazios e “zerados”, esses eram deletados.

Como a parte posterior do Processo de ETL seria realizada no software *Pentaho Kettle*, foram tidas preocupações quanto às linhas “zeradas” de campos numéricos, como o de “Ano de Conclusão”, que não apresentavam valor algum ou apresentavam uma *string* indicando que não havia valor para aquelas linhas, como “-”. Desse modo, tais campos seriam inseridos no *Kettle* como numéricos ao invés de *varchar*, algo que atrapalharia transformações futuras.

## 3 – Tratamento

### 3.1 – Pentaho Kettle

Inicialmente os dados anteriormente trabalhados em forma de planilha no *Microsoft Excel* foram importados em formato .csv por meio do recurso “CSV file input” do *Kettle* com as seguintes configurações:



Após isso, foram criados campos para contabilizar o número de alunos Matriculados, jubilados e com matrícula cancelada por meio da transformação “Value mapper”, conforme os valores presentes no campo “Situação no Curso”.

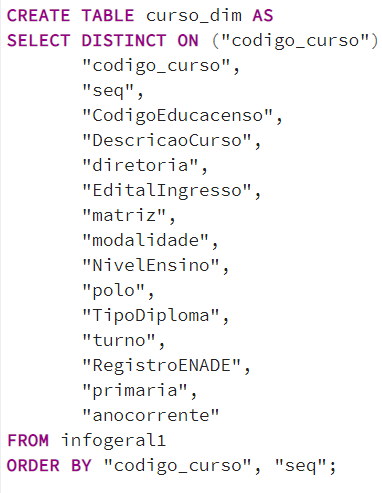
E então foi adicionado um campo com valores sequenciais para servir como chave primária, isso foi feito com a transformação “Add sequence”. Por último foi inserido um campo contendo o ano corrente com a transformação “Add constants”.

Por fim, foram utilizados as ferramenta de transformação e output “Select Values” e “Table output” para mapear os campos e então exportar esses para o banco de dados *PostgreSQL*, o processo realizado no *kettle* pode ser visto a seguir:

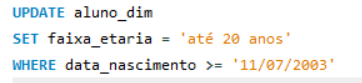


### 3.2 – PostgreSQL

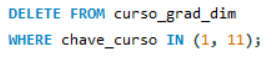
Inicialmente, a tabela importada foi dividida em outras 3, sendo 2 dessas para dimensões e 1 para o fato “situação acadêmica” na forma de transformações como mostrado a seguir:



Então cada tabela foi sujeita à transformações distintas por meio da linguagem SQL, além de haverem mudanças nos nomes de alguns campos para esses serem corretamente manipulados pela linguagem de programação. A tabela “aluno\_dim” teve somente a adição de um campo de faixa etária considerando a data em que a transformação foi realizada, essa foi feita seguinda a forma a apresentada na próxima figura:

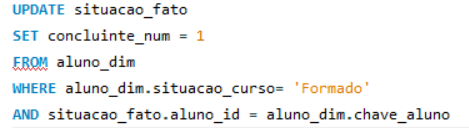


A tabela “curso\_dim” por sua vez teve a ocultação dos dados referentes aos cursos de Biologia e a remoção dos dados dos cursos de pós-graduação, seguindo o formato mostrado:



Além disso, também houve a adição de dados referentes ao período de integralização e número de vagas dos cursos de forma manual.

Por último, a tabela “situação\_fato” foi sujeita à transformações que visavam principalmente a adição de campos numéricos para contabilizar o número de alunos que se encontra em cada situação acadêmica possível, tomando como base os campos criados anteriormente no *Kettle*(ao passo que eles foram criados como varchar) e também os mesmo campos mas considerando somente o período de integralização de cada curso. Além disso, também houve a adição de chaves estrangeiras para referenciar as outras 2 tabelas de dimensão, e a remoção das linhas que faziam referência aos cursos “ocultados” na tabela de cursos. Algumas transformações podem ser vistas a seguir:





## 4 – Implantação na ferramenta de BI

Os dados tratados no PostgreSQL foram então importados para a ferramenta *Microsoft Power BI* para a criação de um ambiente de Business Intelligence. Inicialmente se é estabelecida as relações entre as tabelas fato e dimensão na terceira aba da ferramenta. Os campos para isso foram curso\_id(situacao\_fato) referenciando chave\_curso(curso\_dim) e aluno\_id(situação\_fato) referenciando chave\_aluno(aluno\_dim).



Então foi criada uma tabela para conter métricas adicionais que foram propostas no trabalho. A tabela a seguir apresenta a relação entre as métricas com suas fórmulas no ambiente implementado, é necessário citar que algumas delas referenciam cada curso possível para a realização correta dos cálculos no caso do usuário selecionar mais de 1 curso por vez.

|  |  |
| --- | --- |
| **Métrica** | **Fórmula** |
| Taxa de ocupação inicial | TaxaOcupacaoInicial = SUM('public situacao\_grad\_fato'[ingressante\_num])/SUM('public curso\_dim (2)'[vagas]) |
| Taxa de ocupação inicial efetiva | TxOcupacaoInicialEfetiva =  (      CALCULATE(          SUM('public situacao\_grad\_fato'[ingressante\_efetivo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Agronomia"          )      )      +      CALCULATE(          SUM('public situacao\_grad\_fato'[ingressante\_efetivo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Ciência da Computação"          )      )      +      CALCULATE(          SUM('public situacao\_grad\_fato'[ingressante\_efetivo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Enfermagem"          )      )      +      CALCULATE(          SUM('public situacao\_grad\_fato'[ingressante\_efetivo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Medicina Veterinária"          )      )      +      CALCULATE(          SUM('public situacao\_grad\_fato'[ingressante\_efetivo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Sistemas da Informação"          )      )  )  / DISTINCTCOUNT('public curso\_grad\_dim\_nbio'[descricao\_curso]) |
| Taxa de cancelamento | TxCancelamentoInicial = SUM('public situacao\_grad\_fato'[cancelado\_num])/SUM('public situacao\_grad\_fato'[matriculado\_num]) |
| Taxa de concluintes no prazo | TxConcluintePrazo =  (      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_prazo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Agronomia"          )      ))      +      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_prazo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Ciência da Computação"          )      ))      +      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_prazo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Enfermagem"          )      ))      +      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_prazo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Medicina Veterinária"          )      ))      +      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_prazo\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Sistemas da Informação"          )      ))  )  / DISTINCTCOUNT('public curso\_grad\_dim\_nbio'[descricao\_curso]) |
| Taxa de concluintes | TxConcluintes =  (      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Agronomia"          )      ))      +      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Ciência da Computação"          )      ))      +      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Enfermagem"          )      ))      +      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Medicina Veterinária"          )      ))      +      (CALCULATE(          SUM('public situacao\_grad\_fato'[concluinte\_num]) /          SUM('public curso\_dim (2)'[vagas]),          FILTER(              'public curso\_dim (2)',              'public curso\_dim (2)'[DescricaoCurso] = "Sistemas da Informação"          )      ))  )  / DISTINCTCOUNT('public curso\_grad\_dim\_nbio'[descricao\_curso]) |
| Taxa de concluintes no prazo em relação à ocupação inicial | TxConclusaoPrazo/OcupacaoInicial = SUM('public situacao\_grad\_fato'[concluinte\_prazo\_num]) / SUM('public situacao\_grad\_fato'[ingressante\_num]) |
| Taxa de conclusão no prazo em relação à taxa de conclusão | TxConclusaoTotal = SUM('public situacao\_grad\_fato'[concluinte\_prazo\_num])/SUM('public situacao\_grad\_fato'[concluinte\_num]) |
| Taxa de ocupação média no período | TxOcupacaoMediaPeriodo =  (  (CALCULATE(      SUM('public situacao\_grad\_fato'[matriculado\_periodo\_num]) /      (SUM('public curso\_dim (2)'[vagas]) \* SUM('public curso\_dim (2)'[periodo\_integralizacao])),      FILTER(          'public curso\_dim (2)',          'public curso\_dim (2)'[DescricaoCurso] = "Agronomia"      )  ))  +  (CALCULATE(      SUM('public situacao\_grad\_fato'[matriculado\_periodo\_num]) /      (SUM('public curso\_dim (2)'[vagas]) \* SUM('public curso\_dim (2)'[periodo\_integralizacao])),      FILTER(          'public curso\_dim (2)',          'public curso\_dim (2)'[DescricaoCurso] = "Ciência da Computação"      )  ))  +  (CALCULATE(      SUM('public situacao\_grad\_fato'[matriculado\_periodo\_num]) /      (SUM('public curso\_dim (2)'[vagas]) \* SUM('public curso\_dim (2)'[periodo\_integralizacao])),      FILTER(          'public curso\_dim (2)',          'public curso\_dim (2)'[DescricaoCurso] = "Enfermagem"      )  ))  +  (CALCULATE(      SUM('public situacao\_grad\_fato'[matriculado\_periodo\_num]) /      (SUM('public curso\_dim (2)'[vagas]) \* SUM('public curso\_dim (2)'[periodo\_integralizacao])),      FILTER(          'public curso\_dim (2)',          'public curso\_dim (2)'[DescricaoCurso] = "Medicina Veterinária"      )  ))  +  (CALCULATE(      SUM('public situacao\_grad\_fato'[matriculado\_periodo\_num]) /      (SUM('public curso\_dim (2)'[vagas]) \* SUM('public curso\_dim (2)'[periodo\_integralizacao])),      FILTER(          'public curso\_dim (2)',          'public curso\_dim (2)'[DescricaoCurso] = "Sistemas da Informação"      )  ))  ) / DISTINCTCOUNT('public curso\_dim (2)'[DescricaoCurso]) |

Por fim, o conteúdo foi simplesmente dividido em diferentes páginas conforme sua natureza e foram selecionados visuais distintos conforme a quantidade de “classificações” de cada informação a ser representada, gráficos de pizza e rosca para informações divididas em 2 grupos (sexo, tipo de escola de origem, etc) e gráficos de barra e colunas para informações divididas em mais de 2 grupos (estado de origem, faixa etária, etc). É importante destacar que nos casos de visuais de barra de colunas, quando algumas das “classificações” estavam zeradas, como no caso de não haverem alunos provindos de determinados estados, essas classes eram filtradas para não aparecerem e assim não ocuparem espaço atoa no visual, isso foi feito com o recurso de filtro presente no software *Power BI,* como mostrado a seguir:

