Gerenciamento de informações do IFRN

Nesta etapa do processo de gerenciamento de informações, serão realizadas as seguintes atividades: a integração dos dados do IFRN (todos os campi), a limpeza dos dados, a correspondência dos dados para remover qualquer duplicata, a padronização dos dados, o enriquecimento dos dados e o armazenamento dos dados em um local centralizado.

As atividades a serem realizadas aqui serão as seguintes: as atividades de limpeza e correspondência nos dados, de algumas tabelas na fonte de dados (Sistema Acadêmico) e ou planilhas eletrônicas, com dados obtidos de outros departamentos, como tais setor pedagógico e setor social do IFRN.

A atividade de limpeza identifica dados incorretos ou inválidos, corrige os dados incorretos, propõe correções/sugestões, padroniza os dados e enriquece os dados com mais informações. A atividade de correspondência compara dados e identifica registros semelhantes (mas um pouco diferentes) nos dados que o ajudam a remover duplicatas nos dados.

Para realizar o gerenciamento dos dados do **IFRN**, neste trabalho, estarei utilizando as ferramentas de gerenciamento de dados do *SQL Server 2014*, por dois motivos principais: (1) essas ferramentas estão disponíveis para todos servidores e alunos do **IFRN**, através de convênio assinado entre o IFRN e a Microsoft, (2) essa ferramenta, fornece todos os componentes necessários para uma solução efetiva de Gerenciamento de Informações em um único produto.

Para realizar a limpeza, a correspondência, a padronização e o enriquecimento dos dados e depois, enviar essas informações confiáveis para o armazém de dados (em inglês Data Warehouse) e, para as dimensões da base de dados dimensional (**DDS**), será utilizada ferramenta *SQL Server Data Quality Services* (**DQS)**. Isso é possível, pois O SQL Server Master Data Services (**MDS**) fornece um *hub* de dados central que assegura a constante integridade das informações e consistência dos dados entre diferentes aplicativos.

Essas atividades envolvem os seguintes passos:

1. Criar bases de dados de conhecimento do DQS, para cada tabela a ser analisada.
2. Criar domínios na base de dados de conhecimento para cada tabela a ser analisada, que será usada para a limpeza de dados e correspondência de dados para identificar duplicatas. Uma observação aqui é que, deve-se criar domínios para os campos de dados que se deseja usar nas atividades de limpeza e correspondência, não para todos os campos de dados.
3. Você pode adicionar um valor a um domínio, manualmente ou importando os valores de um arquivo do Excel, executando uma atividade de descoberta de conhecimento nos dados.
4. Você pode defina regras para um domínio. Uma regra de domínio é uma condição usada pelo DQS para validar, corrigir e padronizar valores de domínio. Por exemplo, uma regra poderia ser a seguinte, um E-mail deve ter o formato @ifrn.edu.br.
5. Você pode defina relações baseadas em termos para um domínio. Uma relação baseada em termos, permite ajudar na padronizar os dados e também, na identificação de dados duplicados. Por exemplo, no valor **Aprov. Disciplina e Aproveitamento disciplina.** Que pode ser definido como Aproveitamento disciplina, para todos os registros.
6. Você pode especificar sinônimos em valores de domínio. Você pode definir dois ou mais valores como sinônimos e defini-los como um valor principal, que substitui seus valores de sinônimo durante uma atividade de limpeza para padronizar os dados. Por exemplo, **IFRN** e **Instituto Federal do Rio Grande do Norte** como sinônimos e definir um deles como sendo o principal.
7. E publicar a base de dados de conhecimento para que ela possa ser usada nas atividades de limpeza e correspondência.

A janela da Figura 1 a seguir mostra a tela de gerenciamento do *SQL Server Data Quality Services* (**DQS**), as bases de conhecimentos criadas e seus respectivos projetos de limpeza de dados criados.

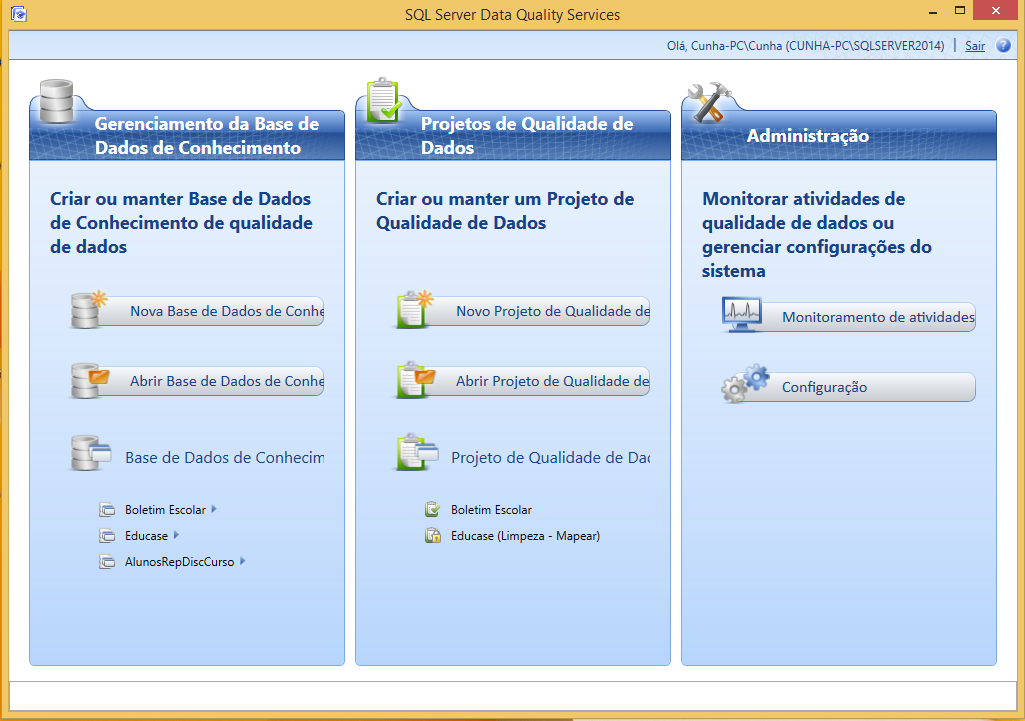


Figura 1 Janela de Gerenciamento de Bases de Conhecimentos do DQS.

Para o gerenciamento das informações da tabela **BoletimEscolar** do Sistema Acadêmico, foi criada uma base de conhecimento e um projeto para limpeza dos dados na ferramenta DQS, como se pode observar na Figura 2. O primeiro passo para se gerenciar um domínio é definir quais atributos se deseja analisar. Para a tabela **BoletimEscolar**, foram definidos os seguintes atributos: campus, matricula, siglaCurso, CoeficienteRendimento, PercentualFreq, Nota, Situacao e turma. O quadro xx mostra um resumo das correções ocorridas em cada atributo analisado.

|  |  |  |
| --- | --- | --- |
| **Atributo** | **Valor origem** | **Corrigido para** |
| Campus | Campu Mossoró | Campus Mossoró |
| Matricula | Não teve correção |  |
| SiglaCurso | Não teve correção |  |
| Situação | Rep. Falta | Reprovado fatal |
| Trans. Curso | Transferência Curso |
| Turma | Não teve correção |  |

Os atributos CoeficienteRendimento, PercentualFreq e Nota, estavam definidos na origem com tipo de dados **string** e foram modificados para o tipo de dados **decimal**. A razão para isso é que, em algumas situações, será preciso realizar cálculos sobre esses atributos.

Após as correções terem sido realizadas, foi feito o **ETL** dessa tabela para a fonte de dados dimensional (**DDSEducacional**).

Gerenciamento de informações da tabela **DadosSociaisAlunos** do Sistema Acadêmico. Também para o processamento da limpeza dos dados dessa tabela, foi utilizada a ferramenta **DQS**. Para a tabela **DadosSociaisAlunos**, foram definidos os seguintes atributos ou domínios: **Situacao**, **Renda\_Familiar**, **Area\_Procedencia\_Escola\_Origem**, **Tipo\_Escola\_Origem**, **Etnia** e **Reside**. Esses domínios foram selecionados, após ter sido feita uma análise na base de dados original (Sistema Acadêmico) e, ter observado alguns valores nulos, para alguns desses atributos e, também por alguns valores, em alguns desses atributos, terem sidos digitados de forma abreviada e outras vezes não. Por exemplo: no atributo **Situacao**, ora tem-se o valor digitado “Aprov. Disciplina” e ora “Aproveitamento disciplina”. Então, nesses casos, foi feita a correção para que todos os valores ficassem com “Aproveitamento disciplina”. A Figura 4 mostra o processo de limpeza dos dados da tabela **DadosSociaisAlunos**, utilizando a ferramenta **DQS**. Para esse caso, no primeiro passo é feito uma descoberta de conhecimento na base de dados, onde se faz um mapeamento entre os atributos da tabela **DadosSociaisAlunos** com os domínios criados no **DQS**. A Figura 5 mostra esse mapeamento.

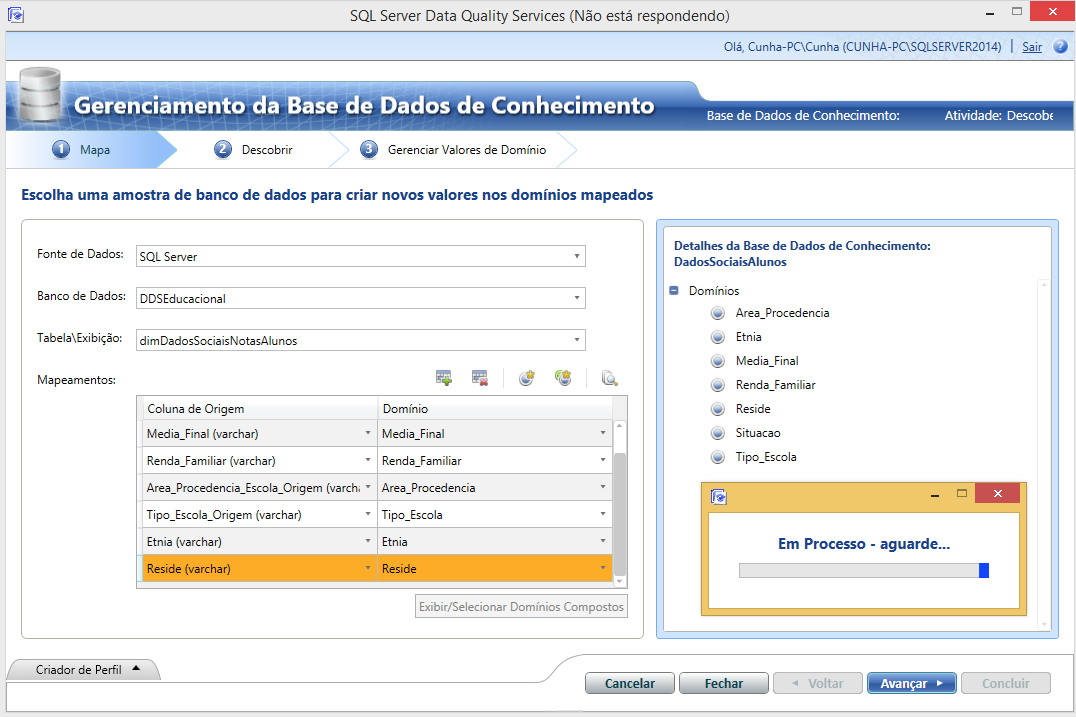


Figura 4 Mapeando atributos e domínios na ferramenta DQS para descoberta de conhecimento.