VALIDAÇÃO DE DADOS EM SISTEMAS DE DATA WAREHOUSE ATRAVÉS DE ÍNDICE DE SIMILARIDADE NO PROCESSO DE ETL E MAPEAMENTO DE TRILHAS DE AUDITORIA UTILIZANDO INDEXAÇÃO ONTOLÓGICA

SANDIR RODRIGUES CAMPOS

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA
UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

VALIDAÇÃO DE DADOS EM SISTEMAS DE DATA WAREHOUSE ATRAVÉS DE ÍNDICE DE SIMILARIDADE NO PROCESSO DE ETL E MAPEAMENTO DE TRILHAS DE AUDITORIA UTILIZANDO INDEXAÇÃO ONTOLÓGICA

SANDIR RODRIGUES CAMPOS

ORIENTADOR: JOÃO PAULO CARVALHO LUSTOSA DA COSTA
COORIENTADOR: RAFAEL TIMÓTEO DE SOUSA JÚNIOR

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPGEE.DM-517/13
BRASÍLIA/DF: JANEIRO - 2013

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

VALIDAÇÃO DE DADOS EM SISTEMAS DE DATA WAREHOUSE ATRAVÉS DE ÍNDICE DE SIMILARIDADE NO PROCESSO DE ETL E MAPEAMENTO DE TRILHAS DE AUDITORIA UTILIZANDO INDEXAÇÃO ONTOLÓGICA

SANDIR RODRIGUES CAMPOS

DISSERTAÇÃO DE MESTRADO ACADÊMICO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

APROVADA POR:

_____
JOÃO PAULO CARVALHO LUSTOSA DA COSTA, Prof. Dr.-Ing. ENE/UnB
(Orientador)

_____
RICARDO ZELENOVSKY, Prof. Dr. ENE/UnB
(Examinador Interno)

_____
EDISON PIGNATON DE FREITAS, Dr. CTEx/EB
(Examinador Externo)

BRASÍLIA/DF, 11 DE JANEIRO DE 2013.

FICHA CATALOGRÁFICA

Campos, Sandir Rodrigues.
VALIDAÇÃO DE DADOS EM SISTEMAS DE DATA WAREHOUSE ATRAVÉS DE ÍNDICE DE SIMILARIDADE NO PROCESSO DE ETL E MAPEAMENTO DE TRILHAS DE AUDITORIA UTILIZANDO INDEXAÇÃO ONTOLÓGICA -- 2013.
xvii, 156 p: il.; 30 cm.

Dissertação (Mestrado) - Universidade de Brasília, Faculdade de Tecnologia. Departamento de Engenharia Elétrica, 2012.
Inclui bibliografia.
Orientação: **JOÃO PAULO CARVALHO LUSTOSA DA COSTA**.

REFERÊNCIA BIBLIOGRÁFICA

CAMPOS, S. R. (2013). Validação de dados em sistemas de *data warehouse* através de índice de similaridade no processo de ETL e mapeamento de trilhas de auditoria utilizando indexação ontológica, Dissertação de Mestrado em Engenharia Elétrica, Publicação, PPGEE.DM-517/13 Departamento de Engenharia Elétrica, Universidade de Brasília, DF, 156p.

CESSÃO DE DIREITOS

AUTOR: Sandir Rodrigues Campos

TÍTULO: validação de dados em sistemas de data warehouse através de índice de similaridade no processo de ETL e mapeamento de trilhas de auditoria utilizando indexação ontológica.

GRAU: Mestre ANO: 2013

Sandir Rodrigues Campos
Qd 02 Conjunto D Casa 121, Setor Norte Gama
CEP: 72.430-204 - Brasília/DF - Brasil

Dedico este trabalho aos meus pais, Sandoval e Nadir.

Sandir Rodrigues Campos

AGRADECIMENTOS

Agradeço a Deus por permitir mais essa conquista.

Aos meus pais Sandoval Ferreira Campos e Nadir Rodrigues Campos pelo voto de confiança no início desta caminhada. As pessoas que estiveram ao meu lado me apoiando mesmo nos momentos que abriram mão da minha presença. Aos meus filhos: Kalel, Davi e Júlia. A minha irmã Aline e seu marido e praticamente irmão Carlos.

Ao amigo Ararigleno Almeida Fernandes pelo apoio incondicional e confiança no meu trabalho e por infiltrar-me no mundo acadêmico.

Aos colegas e amigos da AUDIR do Ministério do Planejamento em especial a Karine pela confiança e por estar ao meu lado.

Aos colegas de pesquisa Toni, Daniel, Clauber e Edison pelo apoio e fundamental ajuda neste momento delicado.

Aos funcionários e amigos do Laboratório de Tecnologias de Tomada de Decisão - LATITUDE. Em especial a Adriana, Andréia e Fábio pela paciência.

Ao meu orientador, colega e amigo João Lustosa. Uma admiração e inspiração. Foi o grande responsável pela minha evolução e por todo o aprendizado.

Ao meu coorientador Rafael Timóteo. Oportunidade e confiança. Os seus conhecimentos foram fundamentais durante todo este processo. Meus sinceros agradecimentos por todas as lições.

Durante o desenvolvimento desta dissertação fui bolsista do CDT/LATITUDE, em projetos de Cooperação Técnica patrocinados pela Secretaria do Patrimônio da União (SPU) e da Auditoria de Recursos Humanos e a Secretaria de Gestão Pública (AUDIR/SEGEP) do Ministério do Planejamento, Orçamento e Gestão.

Obrigado a todos!

"Com grandes poderes, vêm grandes responsabilidades."

Stan Lee

RESUMO

VALIDAÇÃO DE DADOS EM SISTEMAS DE DATA WAREHOUSE ATRAVÉS DE ÍNDICE DE SIMILARIDADE NO PROCESSO DE ETL E MAPEAMENTO DE TRILHAS DE AUDITORIA UTILIZANDO INDEXAÇÃO ONTOLÓGICA

Autor: Sandir Rodrigues Campos

ORIENTADOR: João Paulo Carvalho Lustosa da Costa

COORIENTADOR: Rafael Timóteo de Sousa Júnior

Programa de Pós-graduação em Engenharia Elétrica

Brasília, 11 de janeiro de 2013.

Nesta dissertação se propõe uma estratégia de qualificação de dados em ambiente de Extração, Transformação e Carga (ETL) durante o processo de formação do *data warehouse* (DW).

Como meio de realizar esta qualificação de dados apresenta-se o estudo, a aplicação e a análise dos conceitos envolvidos no processo de validação de dados através da utilização de índices de similaridade. Nesta implementação utilizaram-se duas técnicas: a primeira baseada na adaptação do algoritmo coeficiente de *Dice* e a segunda baseada na distância de Levenshtein.

Além da primeira proposta de qualificação dos dados apresenta-se uma metodologia baseada em mapas conceituais que oferece a incorporação de novas regras de negócio a fim de obter um melhor desempenho na análise de dados. Trata-se das indexações ontológicas de trilhas de auditoria, utilizando como estudo de caso para a sua validação o tratamento dos dados da folha de pagamento dos servidores públicos federais com base no SIAPE.

O processo de incorporação de novas regras de negócio perpassa por toda a criação do mapa ontológico da trilha, até a sua visualização por meio da utilização de uma ferramenta de *Business Intelligence* (BI) onde pode ser observada sua importância para os

procedimentos realizados pela auditoria do Ministério do Planejamento Orçamento e Gestão (MP).

A Ferramenta *open source*, *Pentaho Data Integration* (PDI) integrante da suíte Pentaho *Open Source Business Intelligence*, foi utilizada para implementação das propostas e foram realizados testes funcionais para fins de validação.

Neste estudo foi considerado um ambiente de dados reais que se encontram no arquivo espelho do Sistema Integrado de Administração de Recursos Humanos (SIPAE) e que são disponibilizados no âmbito da parceria celebrada entre o Centro de Desenvolvimento Tecnológico da Universidade de Brasília (CDT/UnB) e a Secretaria de Gestão Pública.

Palavras-chave: Índice de similaridade, ETL, DW, Ontologia, Mapas Conceituais, Qualidade de dados, Folha de Pagamento.

ABSTRACT

ALGORITHMS USING SIMILARITY INDEX BY ETL PROCESS FOR QUALIFICATION DATA SYSTEMS DATA WAREHOUSING

Author: Sandir Rodrigues Campos

Supervisor: Prof. Dr.-Ing. JOÃO PAULO CARVALHO LUSTOSA DA COSTA, Department of Electrical Engineering / University of Brasília

Graduate Program in Electrical Engineering

Brasilia, January 11, 2013.

In this master's thesis, we propose a strategy to data quality in an Extract, Transformation and Load (ETL) environment during the process for creating a Data Warehouse (DW).

In order to acquire this data qualification, we present the study, application and analysis of concepts involving the data validation process by using the similarity indexes. In this implementation, two different techniques are used: the first one based on an adaptation of the Dice coefficient and the second one based on the Levenshtein Distance.

Besides the first contribution on data qualification, we also present a methodology based on concept maps which provides the incorporation of new business rules for obtaining a better performance for the data analysis. This deals on the ontological indexations of the audit trails, using as a case study the payroll information of the public employees stored on the SIAPE database.

The process of incorporation of new business rules goes through the creation of the concept map of the audit trail until its visualization by a Business Intelligence (BI) tool, where the importance of the process for the auditory activities of the Brazilian Ministry of Planning, Budget and Management (MP) can be observed.

The open source tool Pentaho Data Integration (PDI), a part of the Suite Pentaho Open Source Business Intelligence, has been used for the implementation of the proposals. Moreover, functional tests using PDI has been performed for validation.

In this case study, we considered an environment with real data, which is stored in the mirror file of the Integrated System for Human Resources Administration (SIAPE) and which is available due to the cooperation agreement between the Technological Development Centre (CDT) of the University of Brasilia (UnB) and the Brazilian Secretary for the Public Management.

Keywords: Similarity Index, ETL, DW, Ontology, Concept Maps, Data Quality, Payroll.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

BI	Inteligência de Negócios (do inglês Business Intelligence)
DW	Armazém de Dados (do inglês *Data Warehouse*)
ECT	Empresa Brasileira de Correios e Telégrafos
ETL	Extração, Transformação e Carga (do inglês Extraction, Transformation and Loading)
MP	Ministério do Planejamento, Orçamento e Gestão
OLAP	Processamento Analítico em Tempo Real (do inglês Online Analytical Processing)
PDI	Open Source – Pentaho Data Integration
PwC	PricewaterhouseCoopers (empresa multinacional de serviços de contabilidade)
SDW	Sistemas de Data Warehousing
SGBD	Sistema de Gerenciamento de Banco de Dados
SIAPE	Sistema Integrado de Administração de Recursos Humanos
SQL	Structured Query Language
ROI	Return on Investment
SLA	Service Level Agreement (Acordo de Nível de Serviço)
SSD	Sistemas de Suporte à Decisão

1. INTRODUÇÃO

Nos dias atuais, o fator que influi na diferenciação entre as organizações passa por algo muito mais complexo que a introdução ou o uso das tecnologias no desenvolvimento de suas atividades e na produção de seus bens e serviços. Dentro desta realidade atual a inteligência de negócios, do inglês *Business Intelligence* (BI), se apresenta como sendo condição fundamental no auxílio do exercício da tomada de decisões dentro do ambiente das organizações.

Analisando-se este auxílio fundamental no exercício da tomada de decisões a interpretação dos dados gerados pelos diversos sistemas de informação já se tornou uma realidade para as organizações públicas e privadas. Neste intuito, surgiram os Sistemas de Suporte à Decisão (SSD). Dentre estes, os sistemas de *Business Intelligence* foram e estão sendo adotados como parte destas soluções, atuando diretamente em Sistemas de *Data Warehousing* (SDWs).

No domínio de TI, Sistemas de *Data Warehousing* (SDWs) são utilizados como meio armazenamento de grande volume de dados para a aplicação de ferramentas de BI. Os SDWs surgiram para assumir papel de alta relevância nos processos de tomada de decisão constituindo-se numa importante ferramenta de apoio ao processo decisório, atuando como uma plataforma tecnologicamente capaz de disponibilizar um conjunto de meios de se obterem soluções eficazes para as preocupações e necessidades mais essenciais reveladas pelas organizações e em particular, pelos agentes tomadores de decisões [1].

A utilização dos SDWs como ferramenta de auxílio na tomada de decisões se revela importante, na medida em que torna este processo mais rápido e eficaz, flexibilizando o acesso a uma maior quantidade de dados investidos de maior e melhor qualificação, garantindo mais confiança aos responsáveis pelo processo decisório.

Dentro do contexto da utilização dos SDWs é importante salientar a fundamental importância da qualidade dos dados. Para que se possa tirar vantagem dos recursos do Data Warehouse (DW) de forma satisfatória, é preciso que as informações nele armazenadas sejam confiáveis, ou que, pelo menos o grau de confiabilidade das mesmas possa ser considerado durante o processo de tomada de decisão [2].

Ainda dentro desse universo de tomada de decisões se faz necessário o controle da qualidade dos dados manipulados, de maneira a tornar o processo decisório melhor sucedido possível. Logo, torna-se inviável qualquer tomada de decisão baseada em dados inconsistentes.

De maneira geral, a perda de desempenho é associada à ineficácia das ferramentas de interrogação; as arquiteturas de BI são compostas por ferramentas de interrogação e exploração dos dados para geração de relatórios, produzindo informação estratégica para suportar a tomada de decisão [3], ou mesmo a questões de *hardware* relacionadas, em geral, a níveis insuficientes de memória ou a "incapacidade" aparente de processadores. A referida ineficácia pode estar diretamente relacionada à qualidade dos dados, como podemos constatar no estudo da *Price Water House Coopers* (PwC) [4].

Na prática, a importância dos sistemas de DW pode ser realmente atribuída às suas características gerenciais, pois se situam no mais alto plano de ação estratégica das organizações, desvinculando-se do plano operacional e tático. Faz-se importante elucidar os três níveis das organizações, segundo [5], como se verifica na Figura 1.

Figura 1 Níveis Organizacionais

O plano estratégico compreende os altos executivos da organização, responsáveis pela definição dos objetivos da organização e tomada de decisões quanto às questões de longo prazo, tais como: sua sobrevivência, crescimento e eficácia geral [5]. O planejamento, no nível tático, é utilizado para traduzir os objetivos gerais e as estratégias da alta diretoria em objetivos e atividades mais específicos [5]; o principal desafio neste nível é promover um contato eficiente entre o nível estratégico e o nível operacional. Já no planejamento operacional, o processo é de uma menor amplitude, onde o foco é trabalhar junto aos funcionários não administrativos, implementando os planos específicos definidos no planejamento tático. [5].

Os Usuários de SDWs se apresentam como consumidores de informações e conhecimentos, interagindo com o sistema, geralmente por meio de ferramentas que potencializam o processamento analítico dos dados do inglês *On-Line Analytical Processing* (OLAP).

Sempre houve uma preocupação com a melhoria dos processos, porém, a busca por competitividade e qualidade está pressionando os envolvidos na produção a atingirem o

máximo no que realizam. Através de metodologias e técnicas, são implantados programas de qualidade e produtividade para se atingirem as metas [6].

A gestão da qualidade dos dados é uma preocupação assumida desde meados da década de sessenta por parte dos investigadores em estatística [7]. Neste princípio estatístico, a problemática da qualidade de dados se baseava na busca e tratamento dos conjuntos de dados (ex.gr. a duplicação de valores no mesmo conjunto de dados). É, ainda hoje, uma área de ocupação nas investigações no ramo da estatística [8]. Em meados da década de oitenta, as investigações no campo da gestão assumiram a problemática da qualidade dos dados como fonte de pensamento gerador de vantagens estratégicas. Apenas no início da década de noventa, o tema alcançou o devido reconhecimento pela área de tecnologia da informação. Em [9] perspectiva-se a história e regulação da qualidade dos dados de forma similar à ocorrida durante as grandes eras agrícola e industrial.

No intuito de se tornar o entendimento sobre o tema: qualidade, melhor compreendido, este tema foi dividido. Partindo-se deste princípio a década de noventa se dividiu em dois momentos de similar importância.

Num primeiro momento, verifica-se o nascimento da problemática em torno da qualidade dos dados como sendo uma área que se faz merecedora da devida atenção por parte de seus investigadores. Nesta fase surgiram às primeiras investigações a nível acadêmico, quais sejam: em [10]; o programa de TDQM, desenvolvido por Wang [11] e promovido igualmente pelo mesmo instituto; em [12]; o surgimento das primeiras tecnologias de limpeza dos dados, ainda que em fase embrionária, a primeira visão do processo de *extract, transform and load* (ETL), que possuíam o intuito de debelar algumas irregularidades nos dados; Além disso, começaram a surgir algumas conferências de tecnologias da informação abordando o assunto.

Em um segundo momento na década de noventa, verificou-se um aumento significativo dos problemas relacionados com a qualidade dos dados, provocado pelo aumento em qualidade e quantidade das tecnologias de software e pelo aceleramento na perda do

controle dos processos de gestão dos dados [9]. O exponencial crescimento da armazenagem, processamento e compartilhamento de dados, em especial, motivado pelo uso da Internet, evidenciou-se os problemas nos dados existentes e potencializando-se uma nova gama de falhas na qualidade dos dados.

Seguindo esta tendência, os estudos referentes à qualidade dos dados começaram a ganhar corpo e peso dentro do meio acadêmico e organizacional; dessa forma passou a ser encarado como área de estudo independente e, ao mesmo tempo, possuindo um requisito de caráter interdisciplinar.

Motivado pela relevância, diversas investigações, estudos, relatórios e soluções de limpeza de dados implementaram ações em torno da problemática dos dados. A velocidade de criação de tecnologias de software neste ramo de estudos, somada à realização de diversas conferências em nível mundial atraiu muitos investigadores para este tema.

Desta nova corrida em busca da resolução dos problemas encontrados na baixa qualidade dos dados, as organizações, tanto de cunho público como privado, se viram compelidas a possuírem dados de elevado grau de qualidade e desassossegaram-se na ânsia de soluções que alcançassem os seus objetivos [9].

Pós década de 90, verifica-se que o domínio da qualidade de dados atinge um novo patamar, tanto na área de conhecimento como nas tecnologias desenvolvidas com vistas a implantar os conceitos inerentes a esse domínio. As repercussões causadas pela deficiência da qualidade dos dados impulsionaram, as organizações a se enveredarem no sentido de tratar estas questões e a busca de alertas que afiram as reais vantagens em se garantir a melhoria da qualidade dos dados, levam a assumir a qualidade dos dados como uma das prioridades a se resolver [13] [14]. A implementação com sucesso de novas plataformas informáticas, tais como: Sistemas de Data Warehouse (SDW), Enterprise Resource Planning (ERP), Customer Resource Management (CRM) e aplicações Online Analytical Processing (OLAP), não se coaduna com dados de qualidade inferior, até mesmo por seu caráter estratégico, necessitam de dados consistentes, pois são ferramentas de cunho

altamente estratégico e relevante, na tomada de decisões da organização. Ainda no começo do milênio verificam-se iniciativas no campo legislativo no intuito de regular e orientar esta área, traduzindo a importância que este assunto tem assumido nos últimos tempos [14] [15].

O principal intuito na constituição de um DW é a capacidade de prover respostas completas e rápidas aos seus usuários, observando a máxima precisão dos resultados, levando-se em consideração a qualidade da base de dados de origem, o tempo de resposta e a inteligibilidade do resultado apresentado. Para isto, faz-se necessário todo um planejamento baseado na necessidade de apresentação de resultados e informações, de uma maneira que não faltem subsídios importantes e em contrapartida não interfira no desempenho e velocidade das buscas realizadas pelas consultas.

O processo de ETL constitui grande parte dos fundamentos e por vezes o principal fragmento na construção de um DW. A Figura 2 apresenta a representação simplificada do processo de ETL.

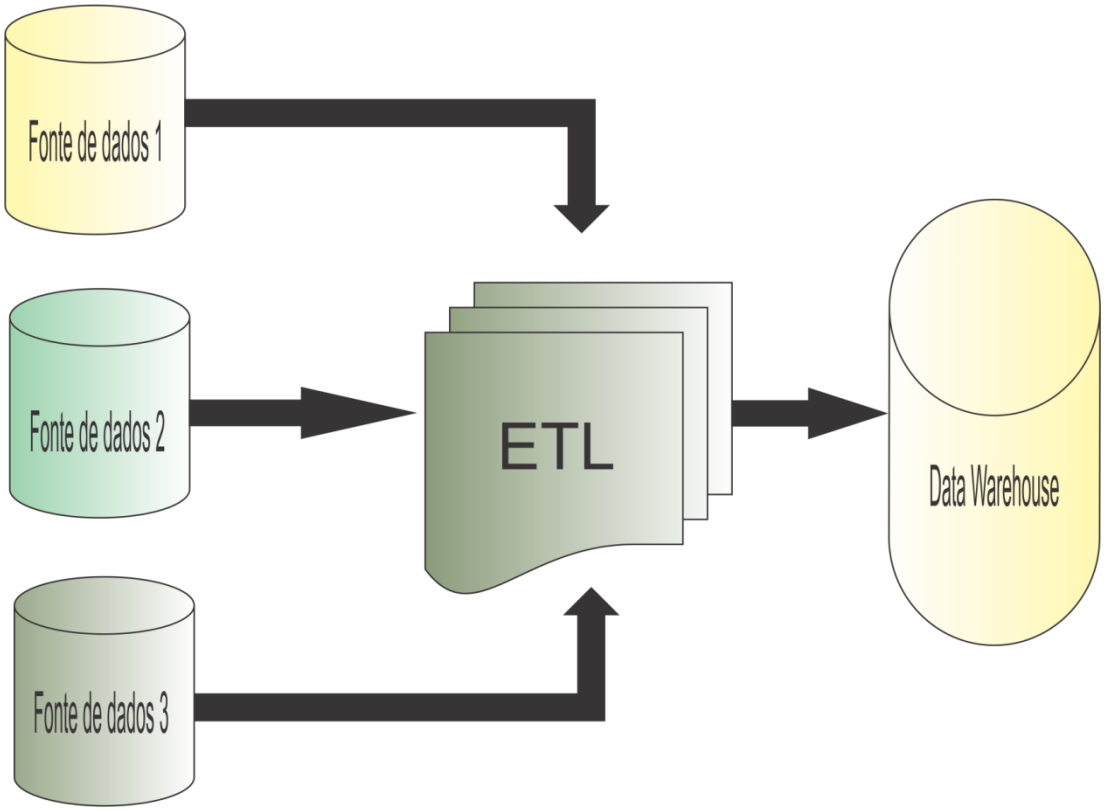


Figura 2 Representação Simplificada do Processo de ETL

O primeiro tema desta dissertação trata de uma técnica que seja capaz de imputar nas tuplas da tabela os atributos de métricas (variáveis) que apresentam valores ausentes ou inconsistentes, utilizando-se de um banco de dados válido com dados oficiais.

Como base validadora dos argumentos, utilizou-se o Diretório Nacional de Endereços (DNE), operando a validação de acordo com número do CEP, comparando o CEP do dicionário de dados com o CEP que se encontra nos arquivos do SIAPE, de maneira a substituir todos os logradouros incorretos ou aqueles com dados ausentes.

Após a utilização desta técnica, é criada uma nova base de dados já se contemplando os novos atributos de enriquecimento, isto é, com todos os atributos que se encontravam ausentes ou inconsistentes, sendo imputados e adicionados a esta nova base de simulação.

A fim de demonstrar a viabilidade prática das técnicas baseadas em índice de similaridade entre *strings* , adaptando-se ao problema em questão, utilizou-se de implementação dentro do processo de ETL. Além disto, este índice é implementado de duas formas: a primeira baseada na adaptação do algoritmo coeficiente de *Dice* e a segunda baseada em lógica *Fuzzy* por meio do algoritmo *Levenshtein Distance*.

O segundo tema desta dissertação trata das indexações ontológicas de trilhas de auditoria, utilizando como caso real a folha de pagamento dos servidores públicos federais com base no SIAPE. O processo perpassa por toda a criação do mapa ontológico da trilha, até a sua visualização por meio de uma ferramenta de *Business Intelligence* (BI) onde pode ser observado, de forma clara, a sua importância para o processo de auditoria do Ministério do Planejamento Orçamento e Gestão (MP).

Para efeito de validação dos dados, faz-se necessária a exata qualificação dos campos imputáveis bem como de suas trilhas de auditoria, principalmente por se tratar de um caso de aplicação real em um projeto de cooperação técnica junto ao MP, ou seja, são dados que serão utilizados para a melhoria dos gastos públicos junto ao SIAPE.

Esta dissertação se utiliza de um caso real que envolve um projeto de pesquisa realizado e pela Universidade de Brasília (UnB) no seu Centro de Apoio ao Desenvolvimento Tecnológico (CDT) e o Ministério do Planejamento Orçamento e Gestão (MP) representado pela sua Secretaria de Gestão Pública (SEGEP), na busca pela melhoria dos gastos públicos, principalmente no que se refere à auditoria do Sistema Integrado de Administração de Recursos Humanos (SIAPE). Para tanto se propõe o desenvolvimento de estratégias de qualificação dos dados extraídos do SIAPE para a criação de um DW consistente e confiável.

1.1. OBJETIVOS E CONTRIBUIÇÕES

A fim de alcançar o objetivo de se proporem soluções para os dois temas principais desta dissertação, que são a utilização de técnicas de índice de similaridade na qualificação do

DW e a indexação ontológica de trilhas de auditoria devem-se resolver os itens intermediários abaixo elencados:

- Demonstrar a importância da qualificação de dados em ambientes de *Data Warehouse* (DW);
- Validar a relevância da indexação ontológica de trilhas de auditoria por meio de mapas conceituais, no intuito desse elaborar uma metodologia de implementação mais eficiente e compreensível;
- Dentro do ambiente de sistemas de DW algumas irregularidades não podem ser resolvidas apenas por meio de simples aplicações tecnológicas; é neste momento que os estudos e investigações tornam-se de fundamental importância e confirmam a forte correlação entre os dados inconsistentes e os elevados custos, despertando nas organizações o devido valor desta problemática;
- O tratamento da informação divulgada como um produto criado pelo sistema de DW, o produto disponibilizado pretende apresentar a informação e maneira que possa ser interpretada pelo gestor do processo decisório;
- A automatização das técnicas de qualificação dos dados ainda no processo de ETL, levando o sistema a corrigir suas próprias inconsistências;

O principal objetivo da presente dissertação consiste em demonstrar por meio de uma aplicação de estudo de caso a aplicação de uma técnica que se utiliza de índices de similaridade para a qualificação do ambiente DW, além disso pretende-se apresentar uma aplicação de indexação ontológica em trilhas de auditoria por meio da criação de mapas conceituais.

Além de resolver os dois temas principais desta dissertação, qualificação dos dados e indexação ontológica, existem objetivos específicos que devem ser alcançados, quais sejam:

- Buscar uma aferição na associação da fraca qualidade dos dados como fator decisivo nos insucessos dos sistemas de DW;

- Definir a qualidade dos dados como sendo um assunto de alta transversalidade dentro da estrutura do DW, atingindo a todos os domínios com diferentes intensidades;
- Encontrar as causas e as consequências da baixa qualidade de dados;
- Compreender, de uma maneira mais ampla, os diversos domínios que abrangem o conceito de qualidade de dados (cliente, produto, produção, excelência e valor);
- Realizar a aplicação de um estudo de caso, sobre bases de dados reais, utilizando-se de algoritmo de pontuação de proximidade para a qualificação de dados em um sistema de DW;
- Mensurar a importância dos sistemas de DW como instrumentos estratégicos das organizações;

Durante o desenvolvimento desta pesquisa, foram aferidos os impactos da utilização de dados inferidos por meio das técnicas de índice de similaridade, verificando-se a obtenção de melhores resultados.

Sugere-se também uma técnica para imputação de dados em um ambiente de DW, a qual possa apontar um caminho para o estudo concernente a este tipo de complementação de dados.

Como maneira de validação desta pesquisa se fez necessário a conferência dos resultados gerados por meio da substituição direta dos valores inconsistentes passíveis de correção durante o processo de ETL, lançando-se mão de um dicionário de dados específico que apontou o caminho a ser seguido pela imputação de dados, baseando-se em índices de similaridades. Aplicando-se as técnicas de imputação e substituição de dados em uma ferramenta que comprove a medida da viabilidade destas técnicas e ainda demonstre sua avaliação. Toda a aplicação do estudo de caso foi baseada na realização de uma comparação entre diferentes metodologias de imputação de dados, aferindo detalhadamente os resultados obtidos.

Durante a realização do trabalho de indexação ontológica das trilhas de auditoria que foi realizado por meio do mapeamento e criação de mapas conceituais realizou-se a sistematização das trilhas de auditoria proposta pelo mapa conceitual lançando-se mão do processo de ETL, para aferição dos resultados obtidos utilizaram-se ferramentas de BI.

1.2. ESTRUTURA DA DISSERTAÇÃO

Esta dissertação é composta de seis capítulos. O segundo intervém diretamente sobre a descrição das tecnologias e metodologias aplicadas durante a elaboração desta dissertação tendo como objetivo principal apresentar a fundamentação teórica de todos os conceitos e aspectos metodológicos abordados durante a elaboração deste trabalho. Suas subseções estão divididas percorrendo os seguintes assuntos: primeiramente descreve-se os aspectos de um *data warehouse* e sua ligação com o *operation data storage*; descrevem-se a seguir a qualidade de dados e sua influência na criação de um SDW bem como suas variadas perspectivas; a seguir descreve-se sobre algoritmos de índice de similaridade, algoritmo de coeficiente de *Dice*, um estudo generalizado sobre lógica *fuzzy* e sua aplicabilidade junto ao algoritmo *levenshtein distance;* a descrição de processo de ETL procura abordar o seu papel durante a elaboração dos dois objetivos desta pesquisa; por último, procura-se estabelecer uma ligação direta entre os conceitos de ontologia e mapas conceituais de maneira a fornecer subsídios para a sua utilização conjunta durante sua aplicação.

O terceiro capítulo está inserido de modo a contextualizar o objeto de estudo e aplicabilidade dos temas propostos nesta dissertação ; Em um primeiro momento procura-se ambientar onde será aplicado o estudo de caso, a seguir identificam-se os problemas pertinentes às soluções apresentadas, perpassando os desafios e os objetivos da proposta. Além disso, apresenta a descrição dos requisitos necessários à realização do projeto, abrangendo todas as especificações das ferramentas utilizadas, bem como toda a preparação para a realização do experimento.

O quarto capítulo trata de descrever a aplicação tecnológica da qualificação de dados durante o processo de ETL, com foco em algoritmo de índice de similaridade baseado em conteúdo, mostrando também a comparação da aplicabilidade desenvolvida com a disponível na ferramenta de BI que se utiliza da lógica *fuzzy*. Com base nesta comparação apresenta-se toda a sua elaboração tecnológica bem como seus resultados obtidos, demonstrando-se a aplicabilidade empírica dos métodos aqui abordados.

No quinto capítulo é descrita a aplicação de ontologia por meio de mapas conceituais no intuito de facilitar o entendimento dos gestores durante o processo de criação de trilhas de auditoria na folha de pagamento dos servidores públicos federais. Essa metodologia acompanha a criação da trilha de auditoria por meio de mapas conceituais, passando pela implementação da trilha criada utilizando-se uma ferramenta de ETL e disponibilizando os resultados obtidos em uma ferramenta de BI de maneira analítica ou por meio de *Dashboards*.

No capítulo sexto são descritas as considerações finais e as principais contribuições e limitações da presente dissertação para a compreensão da problemática relativa à qualidade de dados de um sistema de DW. Além disso, fornece o entendimento dos resultados obtidos nos estudos de casos aplicados e proposições para trabalhos futuros, baseados nos assuntos aqui abordados.

2. FUNDAMENTAÇÃO TEÓRICA

O objetivo deste capítulo está voltado para a apresentação da fundamentação teórica, nele se encontram toda a pesquisa realizada para o desenvolvimento da dissertação e de seus estudos de caso aplicados. No intuito de facilitar a compreensão o capítulo dividiu-se da seguinte forma:

O tema da qualidade de dados é abordado na seção 2.1, de maneira a tornar clara a sua importância no desenvolvimento de sistemas, em particular em ambientes de DW.;

Na seção 2.2 a ideia é abordar de forma sucinta os princípios e fundamentos do *Data Warehouse (DW)* e sua ligação com o *operation data storage*.

Na seção 2.3 faz-se uma abordagem resumida sobre índice de similaridade, seguindo-se, na subseção 2.3.1, de uma descrição sobre o coeficiente de *Dice*.

Na seção 2.4 descreve-se o algoritmo *levenshtein distance*, utilizado para este estudo de caso.

O ambiente de extração, transformação e carga do inglês *extract, transform and load (ETL)* é o tema abordado na seção 2.5.

Como última seção do capítulo que trata da fundamentação teórica desta dissertação temos a seção 2.6 que trata de ontologia e sua ligação com mapas conceituais.

2.1. DATA WAREHOUSE E OPERATION DATA STORAGE

A definição de um DW, segundo [16] seria “um conjunto de banco de dados integrados e baseados em assuntos, onde cada unidade de dados está relacionada a um momento.” De acordo com esta definição, nota-se que um ambiente de DW não consiste apenas de dados. Ele visa integração de variadas fontes e permite um acesso rápido a um número indefinido de dados consolidados por meio de um conjunto de ferramentas para consultar, analisar e apresentar as informações disponíveis [17]. Dessa forma, o DW é indicado na construção de Sistemas de Apoio à Decisão (SAD).

De maneira geral, existe em um DW uma base de dados especializada, resultado da consolidação e do gerenciamento do fluxo de informações oriundos dos bancos de dados corporativos e, inclusive, de fontes de dados externas à organização. Em razão de ter como objetivo principal a análise dos dados, os valores contidos nas bases de um DW têm um foco diferente daqueles inseridos nos bancos de dados operacionais. Esses últimos são os corporativos e estão focados nas operações de um negócio, enquanto os primeiros refletem o histórico das operações e atendem às necessidades dos sistemas de apoio e suporte às decisões gerenciais.

A organização física da base de dados em um DW deve seguir o modelo dimensional (ou esquema estrela) [17]. O esquema estrela é uma estrutura simples, com poucas tabelas e ligações (relacionamentos) bem definidas [18][52]; em um esquema estrela típico de DW, a tabela de fatos contém informações sobre alguma ação que as dimensões realizaram em conjunto [19], assemelha-se ao modelo de negócio, o que facilita a leitura e entendimento, não só pelos analistas, como por usuários finais não familiarizados com estruturas de banco de dados. Permite a criação de um banco de dados que facilita a execução de consultas complexas, podendo ser realizadas de modo eficiente e intuitivo pelo usuário. O nome "estrela" está associado à disposição das tabelas no modelo, que se trata de uma tabela central, a tabela de fatos, que possui relacionamentos com diversas outras tabelas, as tabelas de dimensão. A **Erro! Fonte de referência não encontrada.** apresenta a estrutura eral de um esquema estrela.

Na primeira etapa do processo de ETL (“Extraction, Transformation, and Load”), a etapa da extração, os valores são capturados das múltiplas fontes ou dos múltiplos tipos de uma fonte única, sendo necessárias diferentes ferramentas adaptadas para cada fonte. Tais ferramentas devem ser periodicamente ativadas para capturar do sistema de origem os dados de acordo com a periodicidade programada. Informações de origem e do momento em que um valor surge no contexto de um sistema fonte. A proveniência em base de dados é uma abordagem que permite descrever as informações históricas dos dados, como origem, momento de criação, processos de transformações, entre outros [20]. Por padrão, os fatos em um DW já vêm imbuídos de sua proveniência, por conter o histórico dos fatos.

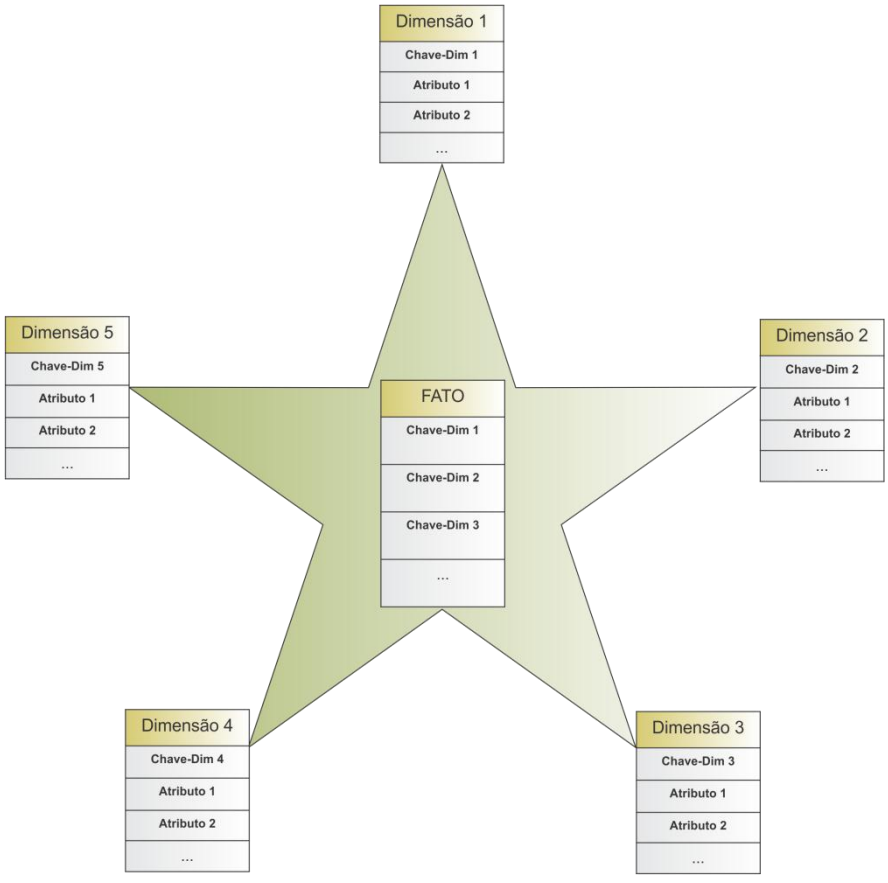


Figura 3 Representação de Modelo Estrela

A **Erro! Fonte de referência não encontrada.** também mostra que não há ligação direta ntre dimensões, todas estão sendo referenciadas diretamente à tabela de fatos. A tabela de fatos é a principal tabela de um modelo dimensional, onde as medições numéricas de

interesse da empresa estão armazenadas [21]. A palavra "fato" representa uma medida dos processos que estamos modelando, como quantidades, valores e indicadores. A tabela de fatos registra os fatos que serão analisados. É composta por uma chave primária (formada por uma combinação única de valores de chaves de dimensão) e pelas métricas de interesse para o negócio. As dimensões indicam a forma como as medidas serão vistas, ou seja, são os aspectos pelos quais se pretende observar as métricas. A intersecção das chaves de dimensão define a granularidade da tabela de fatos, e se torna importante que todas as medidas na tabela de fatos tenham a mesma granularidade.

Frequentemente, os dados corporativos que irão compor as dimensões se encontram armazenados de maneira distribuída, em variadas fontes de dados, ou mesmo sendo composta por diferentes tipos de dados quando da mesma fonte. Pode incorrer em grande variedade de dados, essa variedade pode gerar incoerências (inconsistências) nos valores extraídos das fontes, como erros de digitação, ausência de dados, incoerências entre os metadados e demais distorções geradas por incompatibilidades entre Sistemas de gerenciamento de banco de dados (SGBD).

Assim que detectadas essas incoerências, de maneira a permitir uma análise consistente sobre os valores da base de um DW, faz-se necessária a realização de transformações. Todas as transformações realizadas constituem uma etapa importante no processo de alimentação de um DW, uma vez que uma análise sobre dados não uniformizados pode levar a informações inconsistentes, as quais não refletem a verdadeira realidade de uma corporação e, consequentemente, podem levar a tomada de decisões equivocadas.

Todo o processo de alimentação de um DW não se trata de um processo trivial. Além da etapa de transformação, fazem parte também às etapas de extração e de carga dos dados, ou seja, todo processo de ETL. Usualmente, o processo de ETL é implementado por um conjunto de ferramentas de software, que terão como função a extração da informação de diversas bases e tipos, a transformação dos dados conforme as regras do negócio, e de limpeza e uniformização dos dados, e a carga dos mesmos na base de dados do ambiente de DW [22].

O *Operation Data Storage* (ODS) é um repositório projetado de forma a integrar dados de múltiplas fontes. Além disso, ele se presta a realizar operações complementares sobre os dados ali disponibilizados.

O ODS disponibiliza os seus dados para finalidades operacionais a sistemas e ferramentas que necessitem de sua base de dados, bem como fornece ao DW a fonte para a geração de relatórios gerenciais.

Em razão da multiplicidade dos dados e suas respectivas fontes a criação de um ODS é geralmente precedida pelo processo de ETL onde os dados que serão integrados passam por processos que envolvem a limpeza, a resolução de redundância e verificação de regras de negócio ou dicionário de dados de modo a manter a máxima integridade.

O processo de implementação de um DW tem alguns aspectos que se assemelham ao desenvolvimento tradicional de sistemas, como preparação para as reuniões com usuários, mas possui diferenças que devem ser observadas cuidadosamente [23].

A construção de um ODS é facultativa, entretanto, ajuda em muito a diminuir os esforços de construção de um DW. Todo o esforço de integração entre os sistemas transacionais da empresa seriam depositados no ODS e a carga do Warehouse seria simplificada de maneira incomensurável.

Um ODS é integrado, orientado a assunto, volátil e estrutura tipo *current-valued* (valores atuais), desenhada para atender aos usuários operacionais, em grandes processos de integração, permitindo um melhor desempenho.

Na **Erro! Fonte de referência não encontrada.**, o ODS é visto como uma arquitetura que alimentada por programas de transformação e integração (i/t). Estes programas de transformação e integração podem ser os mesmos programas que alimentam um DW ou programas separados. O ODS, por sua vez, alimenta um DW.

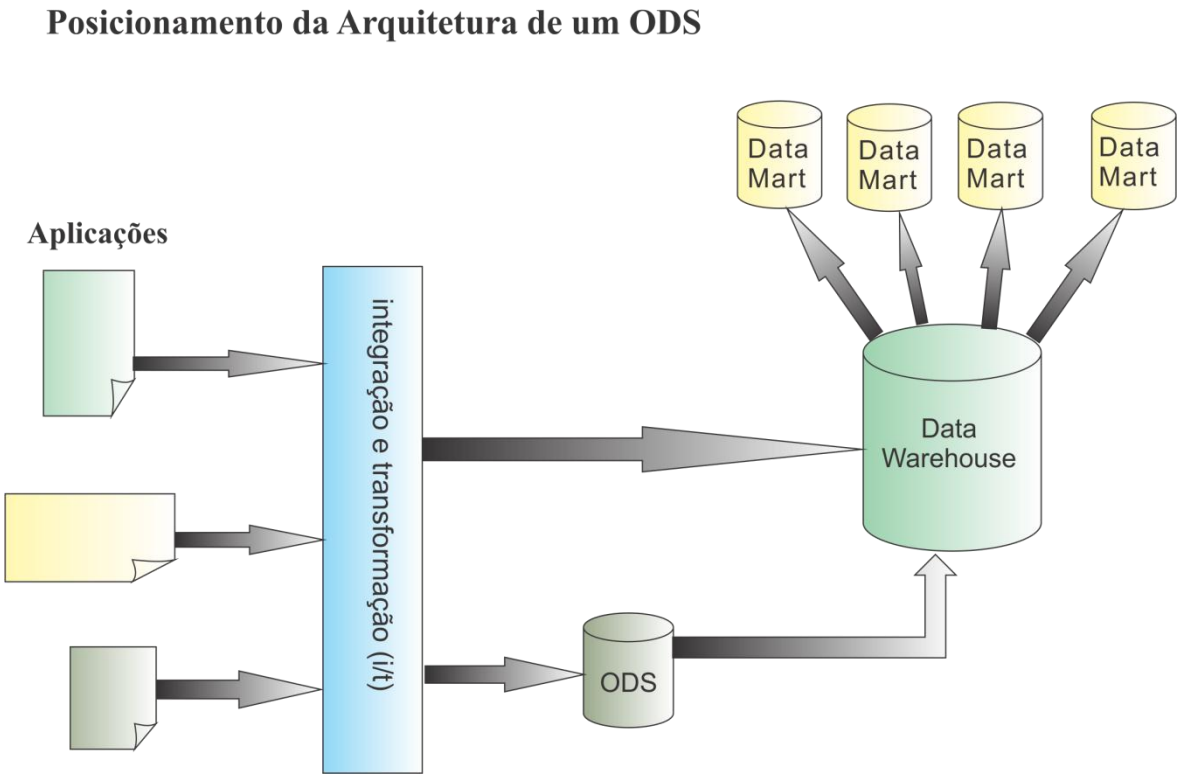


Figura 4 Arquitetura ODS

Alguns dados operacionais podem ir direto para o DW, através da camada de programas de ETL, enquanto outros dados operacionais são enviados para o ODS e depois, do ODS para o *Data Warehouse*.

A essência de um ODS é a de possibilitar um processo on-line de integração coletiva. Um ODS resulta em um desempenho consistente em grandes transações – de dois a três segundos. Um ODS suporta atualizações on-line. Um ODS é integrado com várias aplicações. Um ODS fornece os fundamentos para visões coletivas e atualizadas do negócio. E, ao mesmo tempo, o ODS suporta os processos de apoio à decisão.

2.2. QUALIDADE DE DADOS

No atual contexto organizacional, a informação passou a ter uma fundamental importância face ao ambiente globalizado e altamente competitivo, servindo como grande diferencial na busca de soluções ágeis e com grande probabilidade de acerto. Esta relevância da informação direciona as organizações na busca incessante pela qualificação de suas bases

de dados, pois só desse modo a criação de vantagens competitivas baseadas em informação e conhecimento se torna uma realidade [24].

A experiência acumulada pelas organizações no campo da qualidade dos produtos e serviços permitiu e serviu de suporte para a transferência do conceito ao domínio dos dados ou das informações, conforme se mostra no estudo [25] que considera a qualidade dos produtos ou serviços como dependente dos seus processos de concepção e fabricação. Tal e qual, a qualidade dos dados depende dos processos de planejamento e desenvolvimento inerentes à geração dos dados.

A divisão do tema qualidade de dados é apresentada da seguinte forma: na seção 2.2.1 trata-se dos dados como recursos estratégicos; na subseção 2.2.2 procura-se abordar o contexto geral da qualidade e dados no ambiente de *Data Warehouses* (DW); na subseção 2.2.3 mencionam-se as perspectivas da qualidade de dados; as três perspectivas da qualidade de dados são apresentadas na sequência 2.2.4 a perspectiva ontológica,2.2.5 onde a qualificação dos dados é abordada.

2.2.1. DADOS COMO RECURSOS ESTRATÉGICOS

O domínio das informações por parte das organizações, quer públicas ou quer privadas, é apontado nos dias de hoje como principal diferencial competitivo e fator gerador de recursos econômicos [24] [13]. Neste meio altamente competitivo que apresenta um enorme grau de variabilidade e imprevisibilidade, faz-se necessário a capacitação de se perceber, num primeiro momento, a posição de sua organização perante o mercado e logo a seguir o desenvolvimento de adaptações que gerem vantagens competitivas. Por essa razão a informação assume papel estruturante nas organizações e uma ferramenta de gestão organizacional [26].

Ainda sobre a conceitualização dos dados e maneira a se contextualizar junto ao tema desta dissertação, verificou-se a necessidade de tratarmos com a devida atenção o tema da atualização dos dados, mais especificamente a frequência de suas atualizações. Assim, segundo o estudo [27], podemos identificar três tipos básicos de dados em um ambiente corporativo:

- Dados estáveis: são os dados que possuem mínimas características de volatilidade, considerados de improvável alteração (exemplo de dados no banco SIAPE: nome do servidor, naturalidade do servidor e CPF do servidor);
- Dados que mudam a longo prazo: são os dados que têm uma frequência de mudança muito baixa (exemplo de dados no banco SIAPE: endereço do servidor e escolaridade do servidor);
- Dados que mudam frequentemente: são os dados com alto grau de volatilidade, sujeitos a mudanças intensivas, com uma frequência definida ou de modo aleatório (exemplo de dados no banco SIAPE: rubrica de rendimentos do servidor, salário líquido do servidor).

2.2.2. QUALIDADE DE DADOS EM DWs

O DW é definido, clara e objetivamente em [28], como um sistema de dados orientado por assuntos, integrados, não voláteis e variantes no tempo. Estas características quando devidamente exploradas, afirmam o DW como um meio tecnológico capaz de conceder vantagens estratégicas às organizações no que diz respeito à tomada de decisão, isso posto, fica clara a importância de sua implementação e manutenção dentro do meio organizacional.

O sucesso destes sistemas depende de várias premissas, dentre as quais destacamos: o alinhamento com a estratégia da organização; a engenharia de requisitos adequada; a

solidez e disponibilidade tecnológica que alicerça o DW; as técnicas de modelação utilizadas e a sustentação em processos de extração e atualização dos dados de modo eficaz. A garantia de uma elevada qualidade dos dados, nas mais variadas dimensões, desde as fontes de dados a disponibilização dos dados trabalhados aos utilizadores finais, resume-se como fator primordial para o sucesso destes sistemas [29].

Os sistemas de DW disponibilizam informações, em geral, de uma grande massa de dados no intuito de facilitar o exercício da tomada de decisões, pelos gestores organizacionais, que se utilizam de ferramentas de BI para interpretar estes dados disponibilizados.

Apresenta-se como sendo de fundamental importância a qualidade e integridade disponibilizados no DW. Partindo-se desta análise, fica claro que a qualidade de SDW vai estar diretamente relacionada com a qualidade de seus dados [30]. Dados de má qualidade irão compor um DW de má qualidade, gerando más escolhas no exercício da tomada de decisão.

2.2.3. AS PERSPECTIVAS DA QUALIDADE DE DADOS

Devido ao grande número de vertentes de pesquisas e investigações referentes ao domínio da qualidade de dados, alguns estudos [7] [31] [32] expõem a qualidade dos dados em torno de três perspectivas: a ontológica ou teórica, a intuitiva e a empírica. Em [33] é acrescentada a perspectiva arquitetural.

2.2.4. A PERSPECTIVA ONTOLÓGICA

Dentro da área de Tecnologias de Informação, as ontologias são classificações de terminologias. São usadas como um meio para categorizar ou agrupar as informações em classes. As ontologias também são aplicadas para assimilar e codificar o conhecimento,

definindo as relações existentes entre os conceitos de determinado domínio (uma área do conhecimento) [34].

Dentro desta perspectiva o foco sai dos requisitos dos consumidores e se direciona para o desenho do sistema e a produção dos dados. Assim, é caracterizado um conjunto de métodos e definições, que levam a construção de dimensões da qualidade dos dados [35]. Essas dimensões servem de base para o desenho do sistema de informação e seu modelo ontológico por meio da utilização de mapas conceituais, permitindo a orientação deste no sentido de refletir os aspectos do mundo real [25]. Ao aplicar-se o modelo ontológico por meio da utilização de mapas conceituais pretende-se responder a todos os questionamentos dos gestores de informações.

2.2.5. A QUALIFICAÇÃO DOS DADOS

No estudo apresentado [13], nota-se claramente a evolução como sendo um requisito intrínseco ao termo qualidade de dados. Como ideia original o termo tratava-se apenas como a garantia da exatidão dos dados, porém na evolução de seu entendimento revelou se tratar de um caráter mais amplo e multifacetado. Assim, são consideradas, atualmente outras dimensões, quais sejam, entre outras: a consistência, a completude, a oportunidade e a compreensão.

Com a nova visão multifacetada sobre o termo qualidade de dados, verificou-se um aumento significativo na sua conceituação, até porque mantém um caráter abstrato e subjetivo sobre sua definição, ou seja, não exige a exatidão dos dados, situação impossível e certamente desnecessária [13]. Deste modo, o que deve ser valorizada é a qualidade de utilização da informação gerada. A presença de defeitos é aceitável segundo [49], que considera que os dados não necessitam de estar perfeitos para serem úteis. Este pragmatismo é, igualmente, partilhado em [36], ao ser afirmado que nenhum sistema de informação pode assegurar uma qualidade dos dados de 100%. Partindo desse pressuposto, verifica-se que o problema real não passa por assegurar uma qualidade dos dados perfeita e

sim pela garantia de que o conjunto de informações possua qualidade de dados suficiente e consistente para construção de um DW. Reforçando esta posição, em [37] é considerado como principal objetivo da organização a maximização do valor dos seus recursos a fim de perseguir a missão da organização.

Ainda na mesma investigação, verifica-se que o problema real não consiste em assegurar uma qualidade dos dados perfeita e sim que a qualidade dos dados do sistema de informação seja suficientemente segura, oportuna e consistente, para que uma organização possa sobreviver e tomar decisões razoáveis [36]. Neste sentido, o objetivo da qualidade dos dados consiste em equipar os consumidores com dados, tratados como recurso estratégico, capazes de permitir a inteligência das organizações e possibilitar a tomada de decisões eficazes. Para isso, separa o conceito de qualidade dos dados em duas partes. A primeira, designada por qualidade dos dados inerente e que consiste na representação real dos dados. Uma segunda noção, a qualidade dos dados pragmática é definida como o grau de utilidade e valor dos dados que suportam os processos da organização, visando atingir os objetivos traçados. Neste contexto, os dados armazenados num repositório não possuem valor atual, mas apenas valor potencial.

Portanto, a elevada qualidade dos dados não pode ser entendida como a sua perfeição, essencialmente, por dois fatores. Um primeiro fator remete-nos para uma abordagem multidimensional subjacente ao conceito. Os consumidores dos dados possuem distintas percepções sobre a qualidade dos mesmos dados e um mesmo consumidor pode atribuir importâncias distintas aos mesmos dados em diferentes instantes de tempo. Um segundo fator compreende as razões de natureza econômico-financeira e de ordem logística. A garantia de uma qualidade dos dados perfeita em todas as frentes (dimensões) mostra-se um objetivo tanto improvável como desnecessário. Nesse sentido, importa enveredar os esforços para a obtenção de uma qualidade dos dados adequada a sua finalidade específica, isto é, que satisfaça (ou exceda) os desejos dos consumidores e os auxilie no processo de tomada de decisão.

2.3. ÍNDICE DE SIMILARIDADE

Uma matriz de similaridade é uma matriz de pontuação (escores) que expressa à similaridade entre dois pontos dados. Matrizes de similaridade estão fortemente relacionadas com os seus homólogos, matrizes de distâncias e matrizes de substituição. Os elementos de uma matriz de similaridade medem as semelhanças entre pares de objetos; Quanto maior similaridade de dois objetos, maior o valor da medida.

Nesta dissertação pretende-se utilizar o índice de similaridade na implantação de comparações entre bancos de dados distintos, indicando a maior proximidade de um registro com o seu registro validador.

Atualmente os algoritmos de índice de similaridade que medem distancia podem ser observado em diversos trabalhos como em [38] ou ainda na adaptação do algoritmo de Levenstein Distance [39].

2.3.1. COEFICINTE DE DICE

Entre os algoritmos de índice de similaridade existe o que se baseia no coeficiente de *Dice*, em homenagem a *Lee Raymond Dice.* Este algoritmo busca a medição da similaridade entre conjuntos [40].

O algoritmo definido por *Dice* funciona baseado em pares de caracteres (bigramas). A partir deste fundamento realiza-se a comparação entre *strings* a fórmula matemática aplicada é fundamentada em um coeficiente de similaridade, o valor encontrado varia de 0 a 1, onde zero significaria nenhuma similaridade e um significaria total similaridade o que representa dizer uma igualdade perfeita. Em uma tabela de percentuais 0 seria 0% e 1 seria 100% de similaridade.

Para melhor compreensão exemplifica-se a constituição de bigramas, na palavra "ontologia" os bigramas formariam o seguinte conjunto: {on, nt, to, ol, lo, og, gi, ia}. Como se observa no conjunto de bigramas da palavra ontologia, os bigramas são formados por pares de caracteres de determinada palavra ou *string*.

Na equação abaixo o valor do coeficiente de similaridade é calculado da seguinte forma: o coeficiente de similaridade (s) entre as duas *strings* X e Y é igual a duas vezes o número de bigramas encontrados na intersecção entre o conjunto de bigramas de X e Y. O resultado desta multiplicação deve ser dividido pelo número de bigramas do conjunto X somado ao número de bigramas do conjunto Y.

$$\mathrm{s} = \frac{2|X \cap Y|}{|X| + |Y|}$$

A equação também pode ser expressa da seguinte forma quando tomado como uma medida de similaridade entre *strings*, o coeficiente pode ser calculado entre duas strings, x e y usando bigramas como se segue: [41].

$$s = \frac{2n_t}{n_x + n_y}$$

Nesta equação n_t é o número de caracteres dos bigramas encontrados em ambas as cadeias, n_x é o número de bigramas em cadeia de x e n_y é o número de bigramas em cadeia y.

A seguir apresenta-se um exemplo onde o cálculo da similaridade entre duas *strings*, estas *strings* são tratadas como cadeias de caracteres e separadas em bigramas (conjunto de dois caracteres contínuos).

Realizando comparação entre as palavras *green* e *graen*. Gostaríamos de encontrar o conjunto de bigramas em cada palavra:

- A palavra *green* foi separada por bigrama e se encontrou o seguinte conjunto {gr, re, ee, en};
- A palavra *graen* foi separada por bigrama e se encontrou o seguinte conjunto {gr, ra, ae, en};

Verificou-se a partir deste exemplo dois conjuntos contendo quatro elementos cada um, sendo que a intersecção dos dois conjuntos tem como resultado obtido os elementos (bigramas) "gr" e "en".

Inserindo esses números na fórmula:

$$s = \frac{2n_t}{n_x + n_y}$$

Tem-se:

$n_t = 2$

$n_x = 4$

$n_y = 4$

Aplicando-se os valores tem-se:

$$s = \frac{2(2)}{4 + 4}$$

Obtendo-se como total o valor 0,5, o que corresponde dizer um índice de similaridade de 50%.

2.3.2. ALGORITMO PARA CÁLCULO DA DISTÂNCIA DE LEVENSHTEIN

O algoritmo *Levenshtein Distance* foi criado pelo cientista russo *Vladimir Levenshtein*, em 1965. O foco desta técnica é a avaliação da similaridade entre duas *strings* com base no número de operações necessárias para transformar uma *string* em outra, sendo que as operações possíveis são a inserção, a exclusão e a substituição. A distância zero indica que as *strings* são idênticas. A partir do tamanho de cada *string* é montada uma matriz, onde

serão imputados os custos de cada operação, geralmente, cada uma delas possui custo 1. Ao final das comparações, a distância é dada pela última posição da matriz.

Apesar de ser um algoritmo relativamente antigo, o *Levenshtein Distance* ainda é bastante utilizado atualmente. Na literatura, é comum encontrar a sua aplicação em tarefas como comparação de dialetos, correção ortográfica, reconhecimento da fala, análise de DNA, detecção de plágios, autenticação de assinaturas e vinculação de registros.

A medida de distância de *Levenshtein* será representada pelo número de transformações que uma *string* tem que se submeter para se igualar a outra que está sendo comparada.

Como exemplo de medição *Levenshtein Distance* entre as *strings*: "Rua Independente" e "Rua Independência" a distancia é igual a três, uma vez que sofrerá três interações até que uma *string* se iguale a outra e não há forma de fazê-lo com menos do que três modificações:

1ª interação – Rua Independente → Rua Independence (substituição de "t" por "c")

2ª interação – Rua Independence → Rua Independenci (substituição de "e" para "i")

3ª interação – Rua Independenci → Rua Independencia (inserção de "a" no final).

2.4. AMBIENTE ETL

Ambientes de DW possibilitam a análise de grandes volumes de dados coletados de diversas fontes. De uma maneira geral os valores que são armazenados na base de dados do DW não trazem informações sobre processos de uma única atividade, tendo em vista que o seu principal propósito baseia-se no cruzamento e consolidação de várias bases de informações que tratam de processos distintos. Na implantação do DW como grande

armazém de dados, o intuito passa a ser a estruturação de um repositório de dados responsável por ser a fonte de informações para a tomada de decisão.

Segundo [17] os componentes que formam um DW completo são:

- Sistemas de Origem – sistemas onde se encontram as fontes de dados representam os locais de onde são extraídos todos os valores que irão se integrar à base de dados.
- Data Staging Area – essa é área que cria um ambiente intermediário de armazenamento e processamento de dados oriundos de diversas aplicações e fontes atuando de forma abrangente, desde o acesso à base dos dados nos sistemas de origem até a área de apresentação. Todas as Regras de negócio regem as transformações e organizações realizadas neste contexto. Geralmente, nessa área, os dados apresentam uma granularidade fina, isto é, maiores detalhes possíveis sobre eles.
- Área de Apresentação de Dados – esta área responde pela interface onde o usuário realiza consultas, gera relatórios e outras aplicações gerenciais de análise sobre os dados devidamente organizados.
- Ferramenta de Acesso aos Dados – por meio de ferramentas de interação e visualização, os dados consolidados se tornam acessíveis e visíveis aos usuários.

Um processo ETL concebido de maneira adequada realiza a extração de dados das mais diversas fontes e tipos de valores, dos sistemas que são denominados como fonte de dados *(Data Sources*) executando assim a primeira de sua importante tarefa.

Em um segundo momento a transformação destas informações devidamente extraídas são tratadas sempre no intuito de reforçar a qualidade dos dados, reparar as mais diversas inconsistências, tanto no que diz respeito à falta de valores como na extensa lista de incoerências que podem incorrer neste tipo de processo.

Dentro deste universo o foco deve estar sempre voltado para dar a devida coerência e conformidade, sempre visando a sua adequabilidade às informações associadas à futura apresentação da informação.

Deve-se lembrar de que as fontes distintas de dados, até mesmo por características do processo, devem estar preparadas para serem utilizadas em conjunto, oferecendo aos gestores destas informações a apresentação deste tratamento em formato pronto para que possam tomar decisões.

Todo o carregamento dessas informações é realizado no momento da carga (*Load*) e deve estar preparado para alimentar o DW de maneira organizada e estruturada.

O processo de ETL de uma maneira geral edifica os dados armazenados no ambiente de DW, além disso, realiza esta função de maneira imperceptível para os usuários finais. Segundo [22], embora o processo de ETL seja transparente aos usuários finais, ele consome cerca de 70% dos recursos necessários para a implementação e manutenção de um típico DW.

Sendo assim, o processo de ETL agrega valor significativo aos dados, pois se trata de um conjunto de ferramentas que realiza muito mais do que a agregação de dados pura e simples. Além desta agregação o sistema de ETL se destaca também por:

- Remover e corrigir erros de dados perdidos (dados fora do padrão);
- Fornecer medidas de confiança de maneira documentada e de maneira que possa ser rastreada;
- Realizar a captura de dados transacionais do ambiente OLTP de maneira a serem disponibilizados em séries históricas;
- Ajustar dados de diversas fontes, de maneira a padroniza-los de modo único e disponibiliza-los corretamente em ambiente de visualização (Ambiente OLAP);
- Estruturar e organizar valores de maneira a tornar o ambiente OLAP o mais otimizado possível, influindo diretamente no desempenho de suas consultas;

Neste sentido e de posse de mais conhecimento o processo de ETL, que se apresenta de maneira simples, quando aprofundado se torna um tema de alta complexidade, nesse sentido torna-se fundamental o estudo e exploração exaustiva de tal processo. De maneira geral e até mesmo de uma forma simplista, no imaginário comum se entende a missão comum e fundamental do processo de ETL, como sendo um processo que obtém valores de diversas fontes e de dados e as disponibilizam em modelos que permitam a sua visualização como informação útil.

Este capítulo ainda apresenta a seguinte divisão por seções: na subseção 2.5.1 apresenta-se a divisão do processo de ETL; na subseção 2.5.2 trata-se da extração de dados; na subseção 2.5.3 o tema gira em torno da limpeza dos dados; na subseção 2.5.4 o assunto tratado diz respeito à integração dos dados; na subseção 2.5.5 finaliza-se o tema com a entrega dos dados.

2.4.1. Dividindo o Processo ETL

Segundo [22] existem, basicamente, quatro passos dentro de um processo completo de ETL, a extração, a limpeza, a conciliação e a entrega. Todos esses quatro passos se desenvolvem durante o processo de ETL propriamente dito.

2.4.1.1. Extração de dados

Como primeiro passo tem-se a de extração na qual os dados brutos provenientes de diversos sistemas de origem são usualmente carregados diretamente para a base do DW, ver Figura 5. A partir daí esses dados se apresentam com pouca reestruturação e por vezes são representados de maneira incoerente, ou se sobrepõem ou se contradizem entre múltiplas fontes de dados. Essa ocorrência se dá em razão das múltiplas e diversas fontes de dados que foram desenvolvidas, implantadas e mantidas de maneira independente no intuito de atender a premência específica de suas aplicações, até então, independentes. Por consequência, o resultado de tamanha diversificação apresenta um alto grau de heterogeneidade, ou seja, valores que tratam de um mesmo dado, porém são analisados de

maneiras diferentes causando conflitos computacionais no que se refere a erros sintáticos, estruturais e semânticos. Costumeiramente os dados estruturados são extraídos em formato de arquivos e texto (TXT) ou em bancos relacionais.

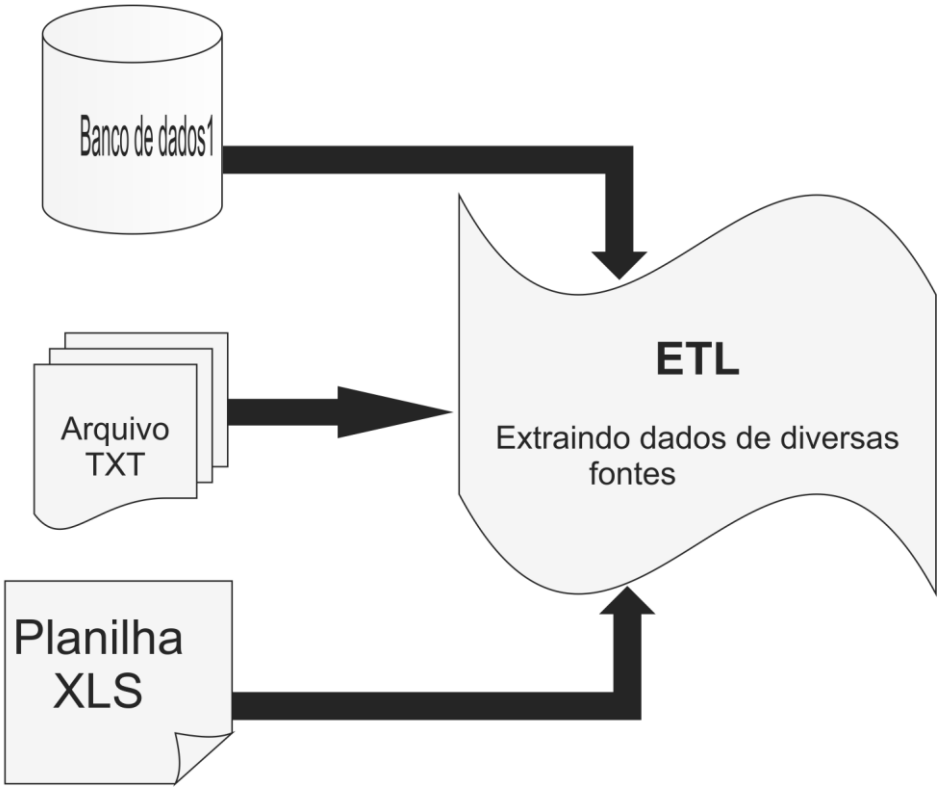


Figura 5 Extração de dados de diferentes fontes de dados

2.4.1.2. Limpeza de dados

Seguindo um passo adiante à extração, os valores, que neste instante estão no *data staging area*, são submetidos ao passo de limpeza, isto é, as transformações dos incoerentes para dados confiáveis (sujos e limpos respectivamente). Como “dados sujos” se entende que são aqueles que apresentam erros ortográficos, redundantes em diferentes representações ou mesmo valores inválidos para determinado tipo de dado. Em [42] é apresentado um estudo mais detalhado sobre os casos de sujeira comuns em bases de dados que o passo de limpeza precisa tratar. Resumidamente, a abordagem de limpeza trata diretamente do aumento da qualidade dos dados.

Segundo os consultores do *Gartner Group*, o processo de "Limpeza de Dados", pode se subdividir nas seguintes etapas, Figura 6:

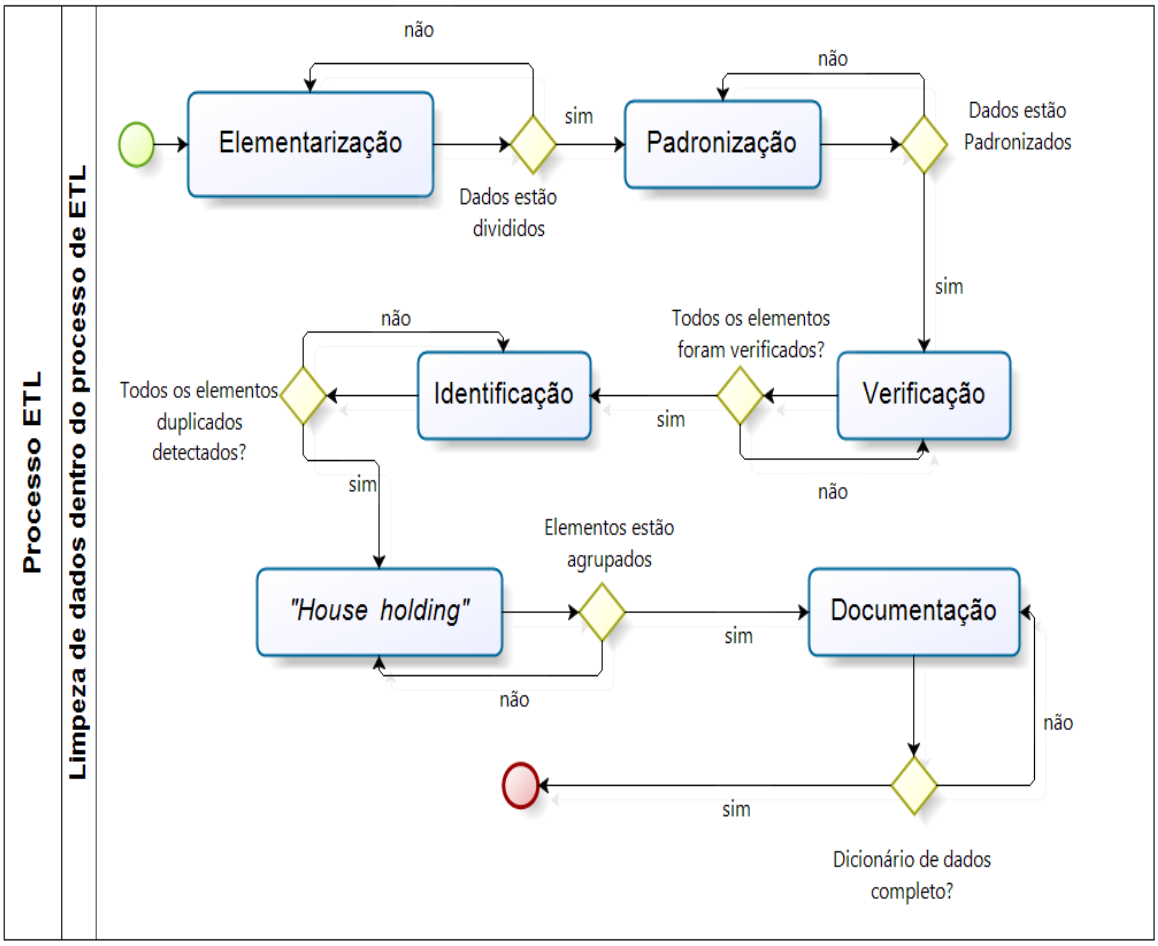


Figura 6 Procedimentos de limpeza de dados dentro do processo de ETL

- Elementarização – dentro do sub processo de limpeza de dados esta etapa distingue-se por infringir a separação dos dados em componentes distintos que receberão a alcunha de "elementos";
- Padronização – nesta etapa se apresenta a aplicação de formato padrão para cada elemento previamente separado;
- Verificação – etapa que varre a base em busca de erros em cada elemento;
- Identificação – momento de detecção de elementos idênticos;
- *House holding* – como o próprio nome sugere trata da economia de dados, identificando grupos de elementos que possuam características em comum;
- Documentação – trata da captura dos resultados dos passos anteriores no intuito da criação de um dicionário de dados que facilite futuros exercícios de limpeza de dados;

É necessário sempre destacar que a limpeza de dados tem o intuito de procurar, encontrar e remover anomalias dos dados no claro interesse de qualificar a base do DW. Usualmente o processo de limpeza de dados não pode ser executado sem o envolvimento de um conhecedor do domínio específico, uma vez que todo o processo de verificação e limpeza de anomalias requer conhecimento especializado sobre os dados a serem tratados. Deste modo, em um primeiro momento podemos considera-lo um processo semiautomático, devendo no decorrer dos diversos processos de aprendizagem ir se tornando o mais automatizado possível, até mesmo em virtude da grande massa de dados que geralmente os envolve e da grande quantidade de recursos aplicados durante o procedimento de limpeza manual.

Uma limpeza de dados que se utiliza de algoritmos deve seguir criteriosamente diversos requisitos. O mais importante destes requisitos consiste na procura e posterior remoção de todas as principais anomalias, quer em níveis de valores individuais ou mesmo no âmbito de todo o conjunto de valores compatível, porém de múltiplas fontes. Os algoritmos implementados para o tratamento por aproximação devem ser genéricos e extensíveis, no intuito de cobrir quaisquer novas fontes inclusas no mesmo processo.

As abordagens que existem para a realização da limpeza de dados é tratada basicamente em dois grandes grupos: especializadas e genéricas.

2.4.1.3. Abordagens Especializadas

Este grupo aborda uma determinada área da limpeza de dados, como no caso da limpeza de endereços postais, ou e problemas mais específicos como nos dados claramente duplicados;

Correção de Nomes e Endereços – A limpeza de nomes e endereços postais tem constante importância nas mais variadas aplicações de processos de ETL. Existem no mercado diversas ferramentas que visam à limpeza deste tipo de dados, estas ferramentas possibilitam a extração e a transformação dos elementos que compõem o nome e o endereço, procedendo à validação de nomes próprios, apelidos, nomes de ruas, localidades e códigos postais, sempre partindo de uma formatação predefinida. Nesta dissertação pretende-se realizar este tipo de aplicação por meio da ferramenta *Open Source Pentaho Data Integration* (PDI) lançando mão de dicionários de dados públicos encontrados na Empresa Brasileira de Correios e Telégrafos (ECT) e na Receita Federal do Brasil e contrapondo-os a base de dados do Sistema Integrado de Administração de Recursos Humanos – SIAPE de maneira a construir um DW mais confiável e com menos inconsistências.

Detecção de Duplicados – Nesta área da limpeza de dados, a busca por dados redundantes, também se encontra uma vasta gama de ferramentas comerciais que se propõem a realizar este tipo específico de correção. Ao utilizar-se a técnica que varre as bases de dados com um algoritmo é importante, para a eficiência do resultado, que cada tupla seja comparada a todas as demais tuplas do processo. Os métodos de aprendizagem de máquinas também podem ser amplamente inseridos nesse contexto a fim de disponibilizar um conhecimento prévio para soluções de ocorrências mais triviais.

2.4.1.4. Abordagens Genéricas

As abordagens genéricas procuram sistematizar um vasto número de inconsistências que devem ser devidamente tratadas durante a etapa de limpeza de dados do processo de ETL. O processo de limpeza de dados recebe um conjunto de fluxos de dados, possivelmente errados e/ou inconsistentes e dai gera um conjunto de fluxos de dados devidamente formatados, corretos e consistentes. Dentro deste escopo algumas ferramentas comerciais de limpeza de dados se utilizam de linguagem declarativa e extensível, geralmente baseada em declarações SQL devidamente qualificadas por um dicionário de dados específicos, que inclusive pode se utilizar de aprendizagem de máquina para acumular experiências neste

dicionário indutivo. Ainda no campo das tecnologias disponíveis cabe à exposição de algumas técnicas utilizadas durante a limpeza de dados no processo de ETL, são elas:

- SQL (*Structured Query Language*) *view* – equivale a uma busca SQL típica, permitindo especificar todo o conjunto de comandos possíveis da linguagem SQL;
- Mapeamento (map) – padroniza o formato dos dados, temos como exemplo dados sobre o sexo em ambientes diferentes de fontes de dados (origem), em alguns ele é definido como masculino e feminino, em outros como F e M, no mapeamento através de intervenções ele pode fundir ou dividir atributos de modo a entrega-los em um formato bem definido;
- Correspondência (match) – procura por pares de tuplas que, com grande probabilidade, referem-se à mesma entidade;
- Segmentação (cluster) – utilizando-se da técnica anterior (correspondência), agrupa os pares de tuplas que possuam alto grau de correspondência, de acordo com específicos critérios de agrupamento, como exemplo de critério podemos citar a transitividade das entidades correspondentes;
- Fusão (merge) – esta técnica realiza a busca em cada segmento de tuplas procurando encontrar e eliminar duplicidades que possam sofrer esta intervenção;

A semântica de cada uma destas transformações envolve a geração de exceções sobre anomalias que possam ocorrer (erros ou inconsistências). Estas exceções são o alicerce de um ambiente que necessita disponibilizar interações para seus usuários. O ambiente também permite a análise dos resultados intermediários, ou seja, em tempo de execução durante o processo de limpeza de dados;

2.4.1.5. Problemas de Qualidade dos Dados versus Limpeza de Dados

No processo de ETL a limpeza de dados cumpre papel fundamental, no entanto é importante notar que aproximações de limpeza de dados podem apresentar alguma inconsistência mesmo depois de realizados seus procedimentos. Dentro deste destaque é

necessário entender que projetos de qualidade de dados demandam um conhecimento específico sobre o ambiente, sobre a criticidade de seus dados e necessitam de preparação e planejamento por parte das organizações. Dentro do estudo realizado sobre a inconsistência em dados postais verificou-se a inexistência de aproximações que permitam efetuar a busca e correção automática de certos problemas de qualidade, tendo em vista o grau de similaridade, ainda se faz necessária a verificação dos dados "limpos". Nestes casos, o tratamento depende da incorporação de uma função extra ou, eventualmente, envolve mesmo a manipulação manual

2.4.1.6. Integração de dados

A Integração de dados trata da estruturação dos dados no esquema de, conciliando-se as dimensões e buscando a padronização da tabela de fatos [43]. A integração deve ser feita para evitarmos que um mesmo elemento em tabelas diferentes tenha nomes distintos. Em SDW's esses dados precisam estar na mesma escala, ou nomenclatura. Na Figura 7 apresenta-se um caso típico de integração de dados. O caso demonstra a integração de dados entre 5 aplicações diferentes, em cada uma dessas aplicações o atributo "Sexo" é tratado de uma forma, cabe a integração de dados definir qual deverá ser o padrão de saída destes dados para o ambiente de DW, além de interpretar cada dado de origem.

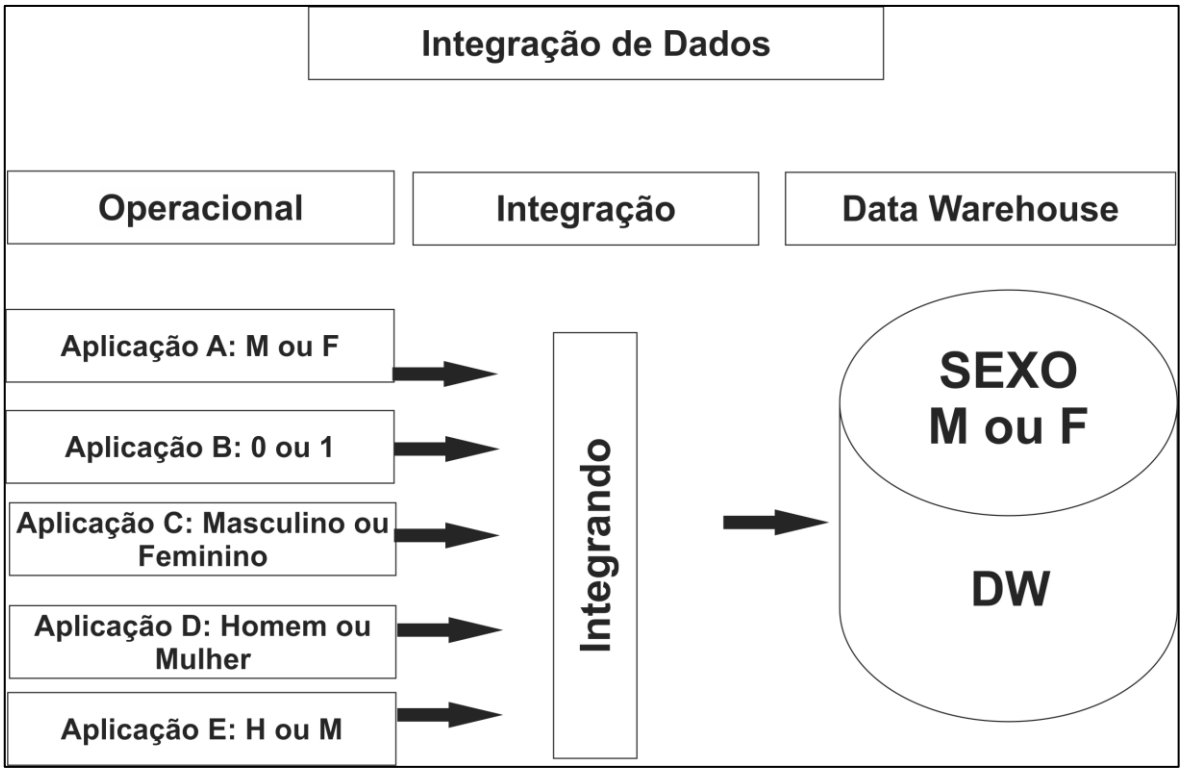


Figura 7 Integração de dados do tipo: SEXO

2.4.1.7. Entrega de dados

Na Entrega de dados sua responsabilidade intrínseca passa a ser a de responder ao processo de ETL, no momento em que os dados estejam prontos para serem multidimensionados para o ambiente. Nesse passo os dados são estruturados fisicamente em conjuntos com esquemas simétricos, conhecidos como modelos dimensionais, conforme Figura 8. Esses esquemas reduzem significativamente o tempo das consultas e simplifica o desenvolvimento de aplicativos [21]. Depois de passar pelo processo de ETL, os dados estão preparados para serem disponibilizados para a camada de apresentação. Todo o processo de ETL convergiu para este momento, no qual os dados são organizados e disponibilizados de maneira útil ao ambiente de DW. Não é regra que o processo de ETL seja sequencial. Suas etapas podem ser executadas em paralelo [22] Em um mesmo momento onde a limpeza de lotes já carregados é realizada a extração de outros lotes de dados pode estar acontecendo ao passo que os dados, lotes "limpos", também já podem ser organizados no esquema proposto para o ambiente de DW. Desse modo o ambiente DW

pode ser operado pelas ferramentas de OLAP realizando combinações entre as dimensões e a tabela de fatos.

No contexto que se chama atualmente de DW 2.0 [44], é buscado, em ambientes de armazéns de dados, uma integração maior entre os dados e seus metadados, além de recuperar e analisar também dados não estruturados, como, por exemplo, e-mails, planilhas eletrônicas, documentos diversos, entre outros. Logo, é preciso que o processo de ETL se adapte a esta realidade.

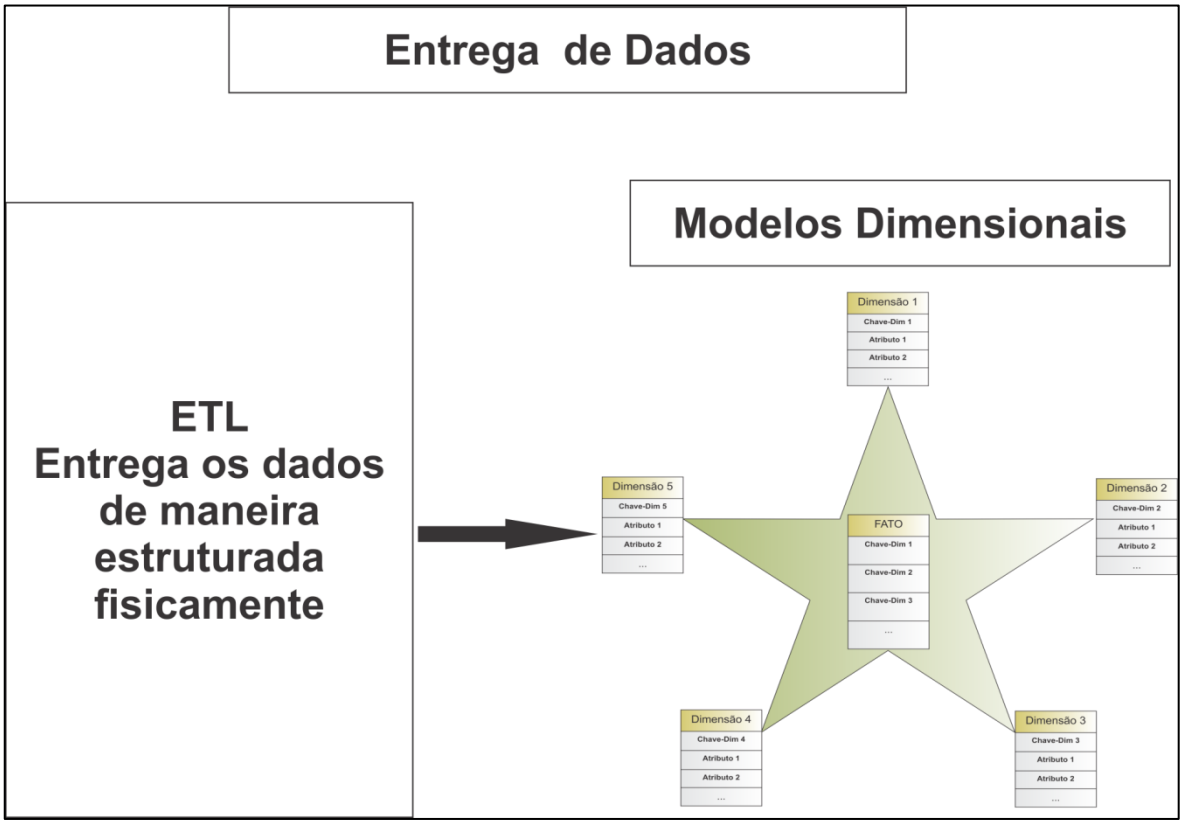


Figura 8 Entrega de Dados

2.5. ONTOLOGIA E MAPAS CONCEITUAIS

No ambiente corporativo, mais especificamente no ambiente da tecnologia da informação nas organizações se torna cada vez mais importante a definição de conceitos a cerca das informações e dados acumulados [34].

Particularmente em ambiente de SDW's onde se lida com grande quantidade de dados, extraídos de diversas fontes as definições quanto aos dados que são tratados pelo processo de ETL e disponibilizados fisicamente em modelos dimensionais as definições claras a cerca das informações se apresentam como sendo de fundamental importância para a geração de um DW confiável que auxilie no processo de tomada de decisão. O estudo que orienta a conceitualização de dados e informações é conhecido como Ontologia [44] e define ontologia como sendo "uma especificação formal e explícita de uma conceitualização compartilhada".

O estudo da ontologia baseia-se fundamentalmente na conceitualização do dado, para realizar esta conceitualização faz-se necessário entender o domínio em que o dado a ser conceituado se encontra.

O domínio irá tratar do ambiente em que o dado se encontra, ou seja, o mesmo dado pode apresentar interpretações diferentes em ambientes distintos, ou mesmo palavras que se apresentam como sendo sinônimas podem apresentar significados diferentes em ambientes específicos. É o caso das palavras provento e salário, que no contexto geral apresentam-se como sendo similares, dentro da ontologia aplicada a trilhas de auditoria possuem significado distinto conforme apresentado em [1].

No mapeamento ontológico as definições são atribuídas a medida que se busca o domínio de origem e por conseguinte a sua definição inicial, uma informação para ser relevante seguiu um caminho para ser formada, utilizando-se de dados e interações, onde termos são compartilhados e por este motivo podem apresentar ambiguidades quando se referem a um mesmo dado com conceitos diferentes.

2.5.1. Construção de ontologias através de mapas conceituais

Em ambientes de SDW's o mapeamento ontológico se apresenta como sendo primordial para o auxilio na construção dos indicadores de desempenho que se busca para o auxilio da tomada de decisões. Por vezes os tomadores de decisão não possuem um entendimento completo de como a informação foi gerada, acarretando em problemas no entendimento das informações a ele disponibilizada [1].

Buscando facilitar o entendimento dos gestores da ontologia de suas informações utiliza-se de mapas conceituais para mapear a ontologia descrita. Segundo [45], um mapa conceitual é uma técnica gráfica para anotar e apresentar conceitos e seus relacionamentos, conforme percebidos por pessoas (individualmente ou em grupo), em relação a um determinado tópico, área de conhecimento, processo, situação. Por meio dos mapas conceituais é possível identificar o conhecimento do gestor de informações que se deseja elicitar [34].

3. CONSTRUÇÃO DO DW DO SIAPE

O objetivo deste capítulo está voltado para a apresentação do contexto em que foram aplicados os estudos de casos que são tratados nesta dissertação. No intuito de facilitar a compreensão o capítulo dividiu-se da seguinte forma:

O tema abordado na seção 3.1, trata de abordar o domínio onde serão aplicados os estudos de caso tratados nesta dissertação.

Na seção 3.2 atribuem-se predefinições da linha de pesquisa bem como da descrição geral do seu meio de aplicação.

Na seção 3.3 faz-se uma abordagem sobre a proposta de aplicação técnica dentro do ambiente de aplicação técnica da pesquisa.

3.1. A COAIS/SEGEP SOB OCONTEXTO DO ESTUDO

Dentre as diversas diretorias e coordenações integrantes da SEGEP, destaque-se a Coordenação de Auditoria de Informações Sistêmicas (COAIS) a quem compete várias atribuições, dentre as quais se se elencam:

- a realização de auditorias na base de dados do Sistema Integrado de Administração de Recursos Humanos (SIAPE), mediante levantamento e análise das rubricas, com vistas a identificar incompatibilidade de vantagens e/ou benefícios, impropriedades, irregularidades e inconsistências cadastrais.
- manter organizado, sistematizado e atualizado os arquivos digitais e físicos, de dados e informações/relatórios da coordenação, bem como manter os registros relativos às apurações sistêmicas nas bases de dados dos sistemas de Governo;
- propor a criação de controles sistêmicos perante a folha de pagamento e aos cadastros do SIAPE;
- elaborar relatórios gerenciais e de auditoria.

Neste ambiente, o projeto se justifica, de maneira empírica, por meio da qualificação da base de dados do SIAPE a ser disponibilizada em ODS, devidamente tratado e qualificado por meio de ferramenta de ETL, de maneira a criar um ambiente de DW para Sistemas de *Business Intelligence* (BI).

Tendo em vista as competências da COAIS, verificou-se a relevante importância do tratamento qualitativo da base de dados, disponibilizado mensalmente, extraído por meio do arquivo fita espelho do SIAPE e demais tabelas auxiliares que compõem a folha de pagamento dos servidores públicos federais.

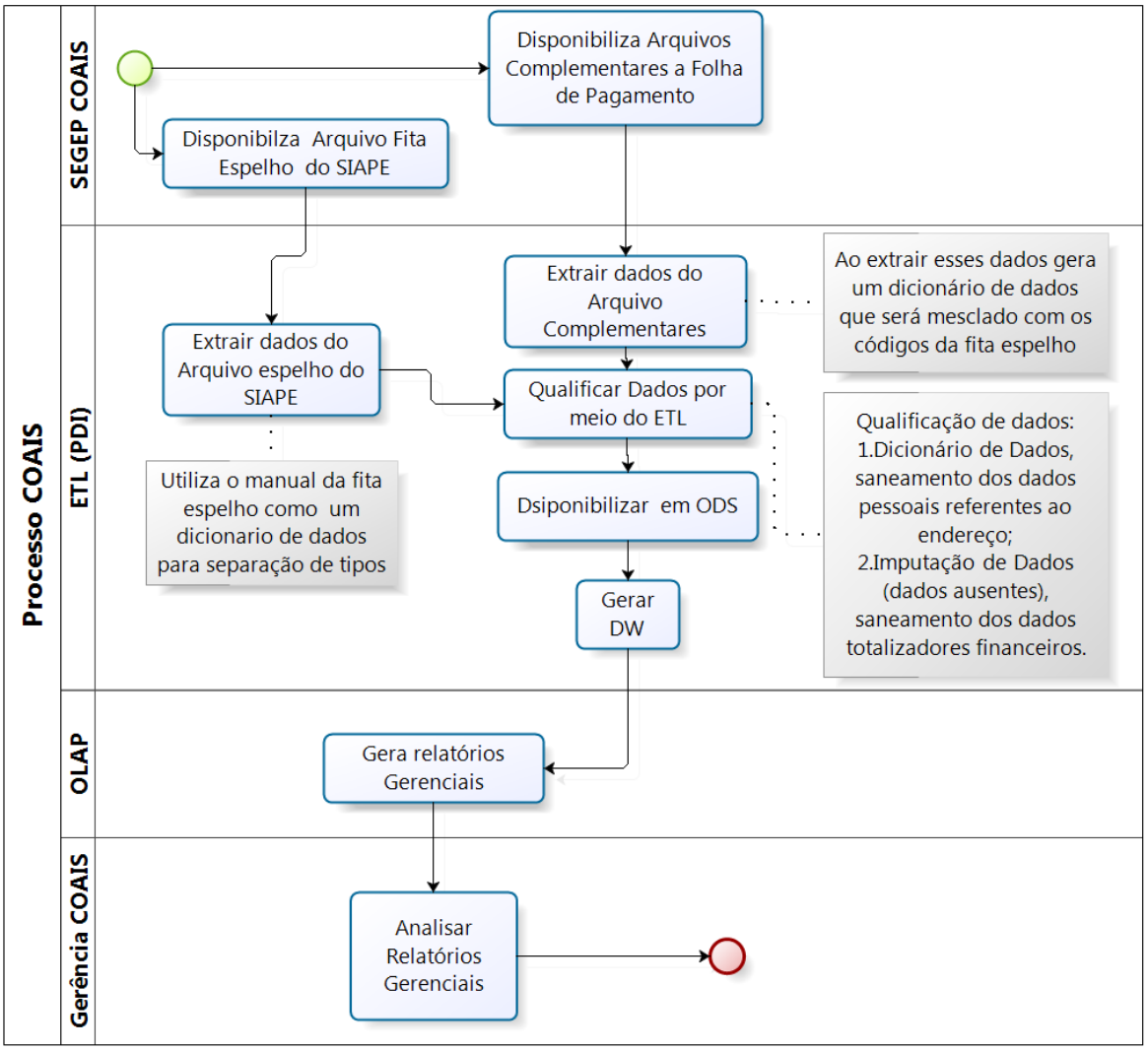


Figura 9: Processo COAIS

Na **Erro! Fonte de referência não encontrada.** observa-se a descrição de todo o rocesso realizado pela COAIS na busca por indícios de irregularidades na folha de pagamento dos servidores públicos federais. Este processo segue várias etapas e pode ser descrito respectivamente da seguinte forma:

1. O processo se inicia com a intervenção do servidor da COAIS que acessa o SIAPE e disponibiliza o arquivo fita espelho do SIAPE e demais arquivos que **compõem** a folha de pagamento;

2. Após o recebimento destes arquivos eles são devidamente extraídos, cada um com sua especificação; o arquivo fita espelho do SIAPE segue um manual, descrito com mais detalhes na seção **3.3,** denominado como Manual da Fita Espelho, este manual descreve exatamente onde se encontra cada

campo no arquivo fita espelho do SIAPE, disponibilizado pelo servidor da COAIS; os arquivos complementares se mesclam ao arquivo fita espelho do SIAPE durante o processo de ETL, trazendo descrições de códigos, totalizadores e demais complementos necessários para a confecção da folha de pagamento dos servidores públicos federais;

3. Após a extração dos arquivos o processo de ETL irá trabalhar na qualificação dos dados; nesta etapa os arquivos são mesclados para a formação do ODS; aqui ocorre a consulta ao dicionário de dados de endereços e a imputação direta de dados nos campos com valores ausentes;

4. Ao término o processo de ETL a ferramenta disponibiliza os resultados em um banco ODS;

5. O banco ODS alimenta o Data Warehouse;

6. A ferramenta de OLAP gera relatórios gerenciais a partir do DW;

7. Os auditores analisam os resultados obtidos por meio dos relatórios gerenciais.

A missão da COAIS/SEGEP/MP depende, em larga medida, da qualidade das informações disponíveis para seus gestores, bem como da garantia e atualização dessas informações. Em especial, os resultados de pesquisas realizadas por meio de trilhas de auditoria ou indícios de irregularidade na folha de pagamento. O projeto foi alçado ao nível de prioridade de programa de governo. No propósito de fornecer informações confiáveis e dados com a integridade necessária se faz extremamente necessária a qualificação dos dados, até como meio de apontamentos de irregularidades cadastrais ou financeiras dentro do banco de dados do SIAPE.

O SIAPE foi criado em 1995, sendo desenvolvido em linguagem Natural, tendo como banco de dados um mainframe OS390 que se utiliza do *software* ADABAS; Além disso o Sistema Integrado de Administração de Recursos Humanos (SIAPE) é a ferramenta responsável por organizar o pagamento de aproximadamente 1,4 milhão de servidores

público federais, possuindo um cadastro completo dos servidores ativos, aposentados e pensionistas.

Com a crescente popularização do uso da INTERNET, surgiram novas oportunidades de distribuição de informações a um número maior de pessoas, com rapidez e segurança, tornando-se viável o desenvolvimento de aplicativos que tratam informações confidenciais e com restrição de acesso. Visando facilitar o acesso dos servidores aos seus dados cadastrais bem como um meio de disponibilização da folha de pagamento dos servidores foi criado o SIAPEnet.

O SIAPEnet é a sigla criada para identificar o sistema de acesso às informações armazenadas nas bases de dados do SIAPE, por intermédio da INTERNET.O SIAPE é um sistema *on-line*, de abrangência nacional, que constitui-se hoje na principal ferramenta para a gestão do pessoal civil do Governo Federal, gerando mensalmente a folha de pagamento de cerca de 1 milhão e 400 mil servidores ativos, aposentados e pensionistas em 214 órgãos da administração pública federal direta, instituições federais de ensino, ex-territórios federais, autarquias, fundações e empresas públicas. De maneira sucinta o SIAPEnet pode ser entendido como sendo um aplicativo que possibilita ao servidor efetuar consultas, atualização e impressão de dados extraídos diretamente do SIAPE.

A fita espelho do SIAPE tem como objetivo gravar em arquivo sequencial os dados Pessoais, Funcionais e Financeiros de Servidores.

O processo de solicitação do arquivo – Fita Espelho do SIAPE é realizado por servidores autorizados da COAIS/SEGEP do MP, junto ao SERPRO, observando-se que a data para solicitação dos arquivos deverá estar de acordo com o cronograma da folha de pagamento SIAPE, mensalmente divulgado a todos os órgãos, pelo gestor do sistema.

Quanto à disponibilidade do arquivo – Fita Espelho do SIAPE, ele estará disponível para download na página do SIAPEnet, após o processamento da folha de pagamento SIAPE e de acordo com o cronograma do sistema, mensalmente divulgado a todos os órgãos, pelo gestor do SIAPE.

Com o apoio da Universidade de Brasília, a SEGEP realizou uma avaliação da qualidade das informações das bases de dados, disponibilizadas por meio da Fita Espelho do SIAPE com o objetivo de identificar problemas de inconsistência, desatualização ou indisponibilidade de dados para a realização dos seus processos.

O trabalho foi realizado por meio de um projeto de cooperação técnica entre a SEGEP/MP e o Centro de Apoio ao Desenvolvimento Tecnológico da Universidade de Brasília (CDT/UnB). Consistiu na realização de um projeto que abrangeu o levantamento e documentação dos processos da área de auditoria, possibilitando a criação de um protótipo que acompanhava todo o processo de maneira automatizada.

A criação do protótipo levou os pesquisadores do CDT/UnB a verificarem a necessidade da qualificação dos dados de entrada, tendo em vista que o arquivo que gerava as trilhas e os indícios de irregularidades, arquivo Fita Espelho do SIAPE, apresenta diversas inconsistências e baixo nível de qualificação de informações.

Aplicou-se, portanto, a metodologia de gestão de qualidade de dados para um levantamento e criação de um Plano de Qualificação das Informações da Fita Espelho do SIAPE.

3.2. REQUISITOS DO PROCESSO DE CONSTRUÇÃO DO DW PARA APLICAÇÃO DO ESTUDO DE CASO

Dentre os grandes desafios encontrados durante a elaboração deste projeto encontra-se a qualificação dos dados em ambiente de DW. Esta qualificação deve ser realizada de maneira tal a disponibilizar informações, levando os auditores a análises corretas no que tange a detecção de inconsistências e tomadas de decisões que produzam correções adequadas aos processos.

A qualificação dos dados em ambiente ETL para alimentação do DW e posterior visualização nos sistemas de BI se torna de alta relevância, pois incorre diretamente nos resultados analisados por auditores responsáveis [1]. Estes auditores irão optar por decisões

que aperfeiçoem seu processo produtivo, tendo como base as informações dos relatórios gerenciais [34].

Baseando-se na qualificação dos dados, por meio de ferramenta de ETL, foi proposta a utilização de duas metodologias diferentes, porém complementares, de modo a qualificar os dados de acordo com a adequação de cada método. O primeiro método consiste na utilização de dicionário de dados dentro da ferramenta de ETL, de maneira a qualificar os dados cadastrais, tornando assim o ambiente de DW mais confiável e verossímil. Já na segunda metodologia é proposto dentro de um ambiente de dados pré-definido a imputação de dados por meio da ferramenta de ETL, lançando-se mão de algoritmos de filtragem de informação.

A Secretaria de Gestão Pública (SEGEP/MP) é órgão da administração direta, vinculado ao Ministério do Planejamento, Orçamento e Gestão, que tem por missão "Promover a excelência da gestão pública na atuação do governo em benefício da sociedade". Este órgão desenvolve desde o ano de 2010 até o ano vigente um projeto de cooperação técnica na área de Inteligência de Negócios (Business Intelligence – BI) e prototipação de uma ferramenta de gestão em parceria com a equipe da Universidade de Brasília - UnB.

Tal projeto vem obtendo sucesso na construção de um modelo de Data Warehouse que armazene informações gerenciais da folha de pagamento dos Servidores Públicos Federais, bem como na prototipação de uma ferramenta que realize a operacionalização dos processos que envolvem as diversas atividades desta secretaria.

Tais resultados fornecem subsídios para realização de auditoria na busca pela excelência da gestão pública, objeto finalístico da SEGEP, obtendo como produtos finais a documentação detalhada que trata a proveniência das trilhas de auditoria a serem meticulosamente especificadas de maneira a permitir a descoberta de indícios de irregularidades.

Como resultado, obteve-se um protótipo de ferramenta que trata de todos os processos mapeados por essa secretaria, bem como, a geração de indicadores gerenciais e seus respectivos mapas conceituais utilizando-se de preceitos de BI.

Mapear todas as inconsistências encontradas durante o projeto e mesmo aquelas que vêm de demandas externas ao próprio órgão é parte integrante ao processo de otimização dos dados, no intuito de os tornarem mais confiáveis. O resultado deste mapeamento aferiu uma grande demanda na qualificação dos dados contidos SIAPE (Sistema Integrado de Recursos Humanos).

A fim de atender a estas demandas, a qualificação de dados implantada em ambiente ETL se apresenta como parte de uma solução viável. Para tanto se faz necessário o conhecimento de todo o ambiente onde serão aplicadas as técnicas de qualificação de dados, desde as fontes de dados, passando pela ferramenta de ETL específica que atuará na qualificação do dado, até a disponibilização em ODS e, por conseguinte, em um DW.

No intuito da criação de um ODS que contenha dados com alta integridade e relevante confiabilidade, procurou-se a aplicação de duas técnicas de qualificação dos dados em ambiente ETL.

A técnica que utiliza algoritmo de similaridade na busca pela qualificação dos dados pessoais do cadastro do SIAPE, utilizando-se como base os dados disponíveis no Diretório Nacional de Endereços (DNE) para qualificação direta dos endereços.

Assim, por se tratar de uma base de dados de extrema relevância para o governo federal, entende-se que todos os esforços no sentido de sanear a base de dados do SIAPE representam ganhos significativos. A estes ganhos se somam as questões de alta prioridade no que tange a excelência dos gastos públicos de forma a dar sentido a este projeto.

Sob tal demanda foi realizado o estudo de caso, considerando o saneamento dos dados do SIAPE e posterior disponibilização em ODS. O software escolhido para implementação do ambiente ETL foi a solução *open Source* PDI (Pentaho Data Integration), software amplamente utilizados no desenvolvimento de projetos de pesquisa vinculados a Universidade de Brasília (UnB) e mais especificamente no Laboratório de Tecnologias da Tomada de Decisão (LATITUDE).

Neste capítulo trataremos de contextualizar os problemas de ausência ou da inconsistência de valores em ambientes de DW, além de trazer ao conhecimento os detalhes e requisitos necessários para aplicação do estudo de caso realizado no ambiente da SEGEP.

3.3. OBJETIVO DA PROPOSTA DE APLICAÇÃO TÉCNICA

Conforme exposto, a SEGEP possui uma grande necessidade de melhoria em seu processo de gestão da qualidade de dados. Neste contexto, o estudo aplicado das metodologias que são utilizadas durante o projeto, buscando a imputação direta para casos de ausência de dados e utilização de um dicionário de dados que possa qualificar um tipo específico de dado, além de preceitos de *Total Data Quality Manager (TDQM)* aplicados às técnicas da ferramenta PDI, na esfera pública torna-se de grande importância.

Em razão da avaliação realizada no decorrer do projeto, verificou-se a necessidade de um serviço de qualificação dos dados extraídos da base do SIAPE.

Verificaram-se as seguintes correções a serem implantadas no ambiente ETL para a geração de um ODS consistente e coerente. Essa verificação conta com serviço de qualidade de dados que deve ser executado de maneira a seguir os preceitos da

metodologia de TDQM e deve seguir as macro etapas dessa metodologia abrangendo os seguintes parâmetros:

Qualificação de ENDEREÇOS:

- Inferência / Correção de CEP, de acordo com o Logradouro informado no cadastro;
- Correção de grafia do Logradouro, através do cruzamento fonético com dicionários oficiais de endereços;
- Validação / Correção de Localidade ou Município;
- Validação / Correção de UF;
- Divisão do endereço em seus componentes básicos: tipo de logradouro, título de logradouro, nome do logradouro, número, complemento, bairro, CEP, cidade, UF.

Qualificação de NOMES:

- Padronização de nomes;
- Eliminação de caracteres especiais e espaços;
- Identificação / eliminação de palavras de baixo calão;
- Divisão do nome em seus componentes básicos: nome composto, primeiro e último nome, nome do meio, pronome de tratamento e grau familiar;
- Inferência de tipo de pessoa – Física ou Jurídica, a partir da análise léxica do nome + documento;
- Inferência de sexo a partir do nome.

Qualificação de TELEFONES:

- Padronização de telefones;
- Divisão em DDD, Prefixo, Sufixo;
- Correção / atualização de Prefixo.

Qualificação de DOCUMENTOS

- Padronização do documento de ordem funcional e Pessoal;
- Validação (dígito verificador) de documentações apresentadas para enriquecimento da base pessoal e funcional;

DEDUPLICAÇÃO

Criação de chaves de comparação (*match codes*), compostas a partir de diferentes combinações de dados (ex: Documento + Nome; Nome + Endereço, etc);

- Marcação de registros duplos;
- Implementação de tabelas "De - Para", ligando todos os registros duplos a um único código na base ODS criada, para posterior retorno dos dados qualificados para os sistemas de produção da SIAPE.

A descrição do arquivo espelho do SIAPE no próximo tópico lista todos os dados que compõem os cadastros do Arquivo Espelho do SIAPE e que formam o escopo mínimo de qualificação de dados a ser considerado no trabalho, servindo de guia para a realização das atividades relacionadas.

Nesta dissertação o estudo de caso foi desenvolvido de maneira a atender a qualificação de dados que envolvem os dados postais do servidor, mais especificamente os dados do CEP e do logradouro, no entanto os demais objetivos na qualificação dos dados do SIAPE ainda constituem uma importante contribuição para trabalhos futuros.

3.4. FUNDAMENTAÇÃO DOS DADOS DE ORIGEM DO SIAPE

Entre todos os requisitos apresentados durante a elaboração e andamento do projeto, entendeu-se que o arquivo fita espelho do SIAPE possui maior grau de relevância.

O arquivo fita espelho do SIAPE apresenta 233 campos de informações que compõem a folha de pagamento, estes campos se encontram divididos em seis grupos de tipos específicos de dados, a saber:

1. O grupo Tipo 0 – *Header* possui 7 campos;
2. O grupo Tipo 1 – Dados Pessoais possui 36 campos;
3. O grupo Tipo 2 – Dados Funcionais possui 153 campos;
4. O grupo Tipo 3 – Dados Financeiros possui 22 campos;
5. O grupo Tipo 4 – Totalização dos Dados Financeiros possui 10 campos;
6. O grupo Tipo 9 – *Trailer*, possui 5 campos;

3.4.1. Objetivo do Arquivo Fita Espelho do SIAPE

O objetivo específico do arquivo fita espelho do SIAPE é o de gravar em arquivo sequencial os dados pessoais, funcionais e financeiros dos servidores públicos federais.

3.4.2. Especificações Técnicas do Arquivo Fita Espelho do SIAPE

O sistema SIAPE foi desenvolvido em mainframe para cadastrar os servidores públicos federais e gerar suas folhas de pagamento. A folha de pagamento é gerada mensalmente e fornece um arquivo denominado arquivo Fita Espelho do SIAPE, que representa os dados pessoais, funcionais e financeiros de cada servidor.

A geração do arquivo Fita Espelho do SIAPE apresenta como resultado um arquivo com extensão "csv" (*comma-separated values* - Valores Separados por Vírgula, é um formato de arquivo que armazena dados tabelados, cujo grande uso data da época dos mainframes; por serem bastante simples, arquivos .csv são comuns em todas as plataformas de computador; o CSV é uma implementação particular de arquivos de texto separados por um delimitador, que usa a vírgula e a quebra de linha para separar os valores; O formato também usa as aspas em campos no qual são usados os caracteres reservados).

Este arquivo geralmente possui tamanho variável, pois depende de vários fatores que influem diretamente em seu tamanho, até porque trata da situação funcional de, aproximadamente, um milhão e quatrocentos mil servidores públicos federais, entre ativos, inativos e instituidores de pensão. Atualmente, apesar deste caráter de variabilidade no tamanho do arquivo espelho ele tem se mantido em aproximadamente 13, dado o tamanho da base de dados que o constitui.

No arquivo fita espelho do SIAPE, ou seja, no arquivo CSV gerado, existe uma posição (18), em cada registro (linha de dados contínuos), que define de qual tipo será o registro, ou seja, a ferramenta de ETL ao realizar a separação deste arquivo por tipo deverá verificar em cada registro a posição 18 para classifica-lo, como visualizado na Figura 10.

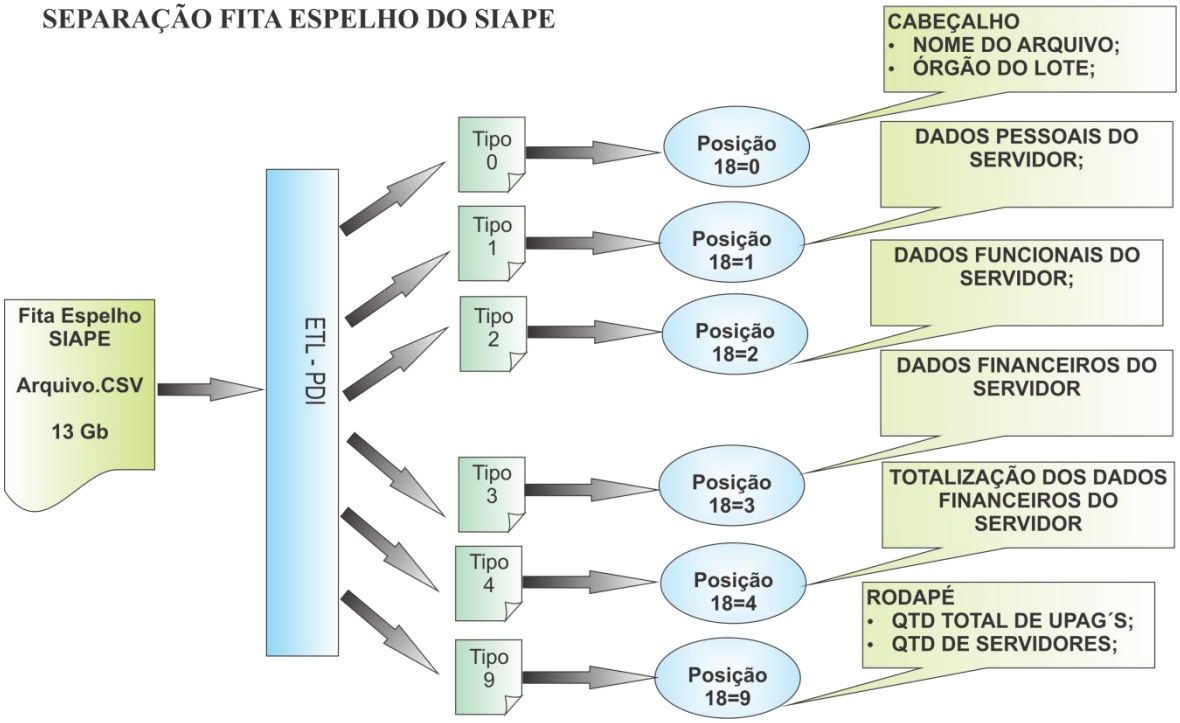


Figura 10 Separando o arquivo fita espelho do SIAPE

3.4.3. Classificação e Organização do Arquivo Fita Espelho do SIAPE

O arquivo está classificado em ordem ascendente pelas posições 1 a 27 de cada registro.

A sua organização física esta descrita na tabela **2** que resume a sequencia de registros.

Tabela 1 Tipos de registros do SIAPE

SE O REGISTRO É DO TIPO	O REGISTRO SEGUINTE SERÁ TIPO
0	1
1	2
2	1,3 ou 9
3	3,4 0u 0

SE O REGISTRO É DO TIPO	O REGISTRO SEGUINTE SERÁ TIPO
4	1 ou 9
9	Fim de Arquivo

Além destas especificações é importante mencionar o formato e alinhamento dos campos.

O campo numérico possui alinhamento à direita e é completado com zeros à esquerda. Quando referente a valores financeiros, os dois últimos dígitos representam as casas decimais dos centavos.

Para os campos Alfas e Alfanuméricos o alinhamento será registrado pela esquerda e completado com brancos à direita.

3.5. FUNDAMENTAÇÃO DOS DADOS CONFRONTANTES – DIRETÓRIO NACIONAL DE ENDEREÇOS (DNE)

O e-DNE, é um banco de dados (não é um programa ou software. não há adoção de software complementar, sendo o desenvolvimento do mesmo a cargo do cliente/usuário) que contém mais de 900 mil CEP de todo o Brasil, constituído de elementos de endereçamento (descrição de logradouros, bairros, municípios, vilas, povoados) e Códigos de Endereçamento Postal - CEP.

É a base oficial e exclusiva dos Correios, sendo assim, a informação é confiável e atualizada. Pode ser comprado em poucos minutos pela Internet, na loja virtual dos Correios, sem a necessidade de formalização de contrato, apenas por meio de adesão ao

Termo de Compromisso. Na compra do e-DNE, as 3 primeiras atualizações são gratuitas, com validade de 01 ano. Melhora a qualidade do seu banco de dados e maximiza a sua comunicação com seus clientes;

3.5.1. Características Gerais

Quem pode usar:

- Pessoas físicas e jurídicas;

Formato dos arquivos contidos no e-DNE:

Dependendo da modalidade, os arquivos serão disponibilizados nos formatos MS-Access (. mdb) e texto (.txt), ou apenas texto (.txt).

Conteúdo do e-DNE:

- Nomes oficiais das ruas de todas as capitais do país e de mais 320 cidades que possuem mais de 50 mil habitantes;

- CEPs dos 5.565 municípios do país;

- CEPs de Distritos e Povoados;

- CEPs das Unidades dos Correios;

- CEPs de Grandes Usuários;

- CEPs das Caixa Postais Comunitárias;

- CEPs promocionais.

Periodicidade de atualização dos dados do e-DNE:

Trimestral.

Na aquisição de qualquer modalidade, as 03 primeiras atualizações já estão inclusas sem custo adicional. (Validade de 1 ano)

Após esse período, para manter a Base de Dados de Endereçamento atualizada, não será necessária a aquisição de outra Base do e-DNE, apenas da Base de Atualização, seja e-DNE Básico ou e-DNE Master.

Modalidades:

e-DNE BÁSICO

É destinado a qualquer pessoa física ou jurídica que:

a) na condição de usuária final trabalhe com as seguintes aplicações:

- Validação, higienização e duplicação de registros de cadastros de endereços;

- Mineração de dados (data mining);

- Localização geográfica (geomarketing);

- Atendimento *call centers* e/ou *telemarketing*;

- Acesso à internet ou intranet;

- Gerenciamento da relação com clientes;

- Customer Relationship Management (CRM);

- Captação de endereços;

- Enterprise Resource Planning (ERP).

Formato dos arquivos contidos no e-DNE BÁSICO: Arquivos formato texto (.txt)

e-DNE MASTER

É destinado a qualquer pessoa física ou jurídica que:

a) na condição de usuária intermediária trabalhe com as seguintes aplicações:

- Desenvolvimento de software para tratamento de endereços e comercialização no mercado corporativo;

- Prestação de serviços de higienização de cadastros de endereços (one-shot ou online);

- Publicação de guias de endereços, catálogos telefônicos e assemelhados

b) na condição de usuária final trabalhe com as seguintes aplicações:

- Validação, higienização e duplicação de registros de cadastros de endereços;

- Mineração de dados (data mining);

- Localização geográfica (geomarketing);

- Atendimento *call centers* e/ou *telemarketing*;

- Acesso à internet ou intranet;

- Gerenciamento da relação com clientes;

- Customer Relationship Management (CRM);

- Captação de endereços;

- Enterprise Resource Planning (ERP).

Formato dos arquivos contidos no e-DNE MASTER: Arquivos formato texto (.txt) e MS-Access (.mdb)

3.5.2. Benefícios

Com a utilização correta dos dados de endereçamento possibilita maior efetividade na distribuição de suas comunicações:

a. Aumento do retorno esperado para as suas ações de marketing;

b. Diminuição das cartas/malas diretas devolvidas por problemas de endereçamento;

c. Redução de custos com envios de cartas/malas diretas para endereços com dados inexistentes/incorretos.

4. VALIDAÇÃO DE DADOS EM SISTEMAS DE DATA WAREHOUSE ATRAVÉS DE ÍNDICE DE SIMILARIDADE NO PROCESSO DE ETL

O objetivo deste capítulo está voltado para a apresentação do ambiente de aplicação e de suas necessidades que originaram a pesquisa empreendida nesta dissertação. No intuito de facilitar a compreensão o capítulo dividiu-se da seguinte forma:

O tema abordado na seção 4.1, trata da descrição da aplicação e sua utilização com meio de solucionar as demandas exigidas durante sua implementação.

Na seção 4.2 faz-se a apresentação dos resultados obtidos por meio da utilização das duas técnicas aplicadas, na subseção 4.2.1, apresenta-se o resultado obtido por meio da distancia de Levensthein e na subseção 4.2.2 os resultados apresentados são os que foram obtidos pelo método de coeficiente de *Dice*.

4.1. DESCRIÇÃO DA APLICAÇÃO DO ESTUDO DE CASO – MÓDULO DE QUALIFICAÇÃO DE DADOS POR MEIO DA UTILIZAÇÃO DO PDI

Para efeito do estudo de caso aplicou-se condições específicas para realização das operações por meio do PDI, é importante notar que a pesquisa foi aplicada a três tipos de consultas para a qualificação de dados postais contidos no SIAPE.

Faz-se necessário elucidar que o trabalho constitui-se na leitura do banco de dados gerado pelo tratamento do arquivo espelho do SIAPE em conjunto com os dados oriundos do DNE, por meio da ferramenta PDI, depois os dados são comparados, segundo os critérios estabelecidos abaixo, por meio de duas metodologias diferentes a fim de aferir qual seria a melhor solução a ser aplicada neste caso específico.

Os critérios definidos para a execução desta aplicação são pela ordem:

1. Encontrar na base SIAPE os servidores que possuem em seu cadastro funcional a seguinte inconsistência em seu cadastro de endereço: não possui número de CEP e concomitantemente possui logradouro – encontrar o CEP correspondente;
2. Encontrar na base SIAPE os servidores que possuem em seu cadastro funcional a seguinte inconsistência em seu cadastro de endereço: possui número de CEP e concomitantemente não possui logradouro - encontrar o logradouro
3. Encontrar na base SIAPE os servidores que possuem em seu cadastro funcional a seguinte inconsistência em seu cadastro de endereço: possui número de CEP "genérico" (possui todos os 3 últimos dígitos sendo igual a zero) e concomitantemente possui logradouro – detalhar o número do CEP;

A ferramenta de ETL traz em sua suíte de opções uma implementação que se utiliza de lógica *fuzzy*, mais especificamente sobre o algoritmo de levenshtein distance, que atribui

pontuações em comparações de cadeias de caracteres, na busca pela qualificação de dados específicos.

Além da opção padrão disponibilizada pela ferramenta de ETL, buscou-se uma adaptação de um algoritmo que trabalhasse com índice de similaridade para realizar uma comparação entre resultados, tendo em vista que a opção padrão da ferramenta de ETL poderia não apresentar os melhores valores para o estudo de caso específico. Para tanto utilizou-se uma adaptação do coeficiente de Dice para desenvolver uma aplicação que pudesse ser inserida no processo de ETL para qualificação dos mesmos dados.

4.2. APRESENTAÇÃO DAS DOS RESULTADOS OBTIDOS

Os resultados serão apresentados em duas etapas, a primeira etapa traz os resultados obtidos por meio da distancia de *Levenshtein*, já a segunda etapa trará os resultados obtidos por meio do coeficiente de *Dice*.

4.2.1. Resultados obtidos por meio da distancia de Levenshtein

A Figura 11 traz a tabela analítica dos resultados obtidos por meio da distancia de *Levenshtein* e o gráfico de barra resultante. O gráfico traz as dimensões de distância de 0 a 10 e ainda os resultados nos quais não foi possível encontrar similaridade com o resultado

igual a *null*. Os resultados obtidos no passo três sob a técnica de distancia de *Levenshtein* foram limitados a mil (1000) registros devido ao alcance de tempo excedido ao tentar-se implementar a técnica em todos os registros.

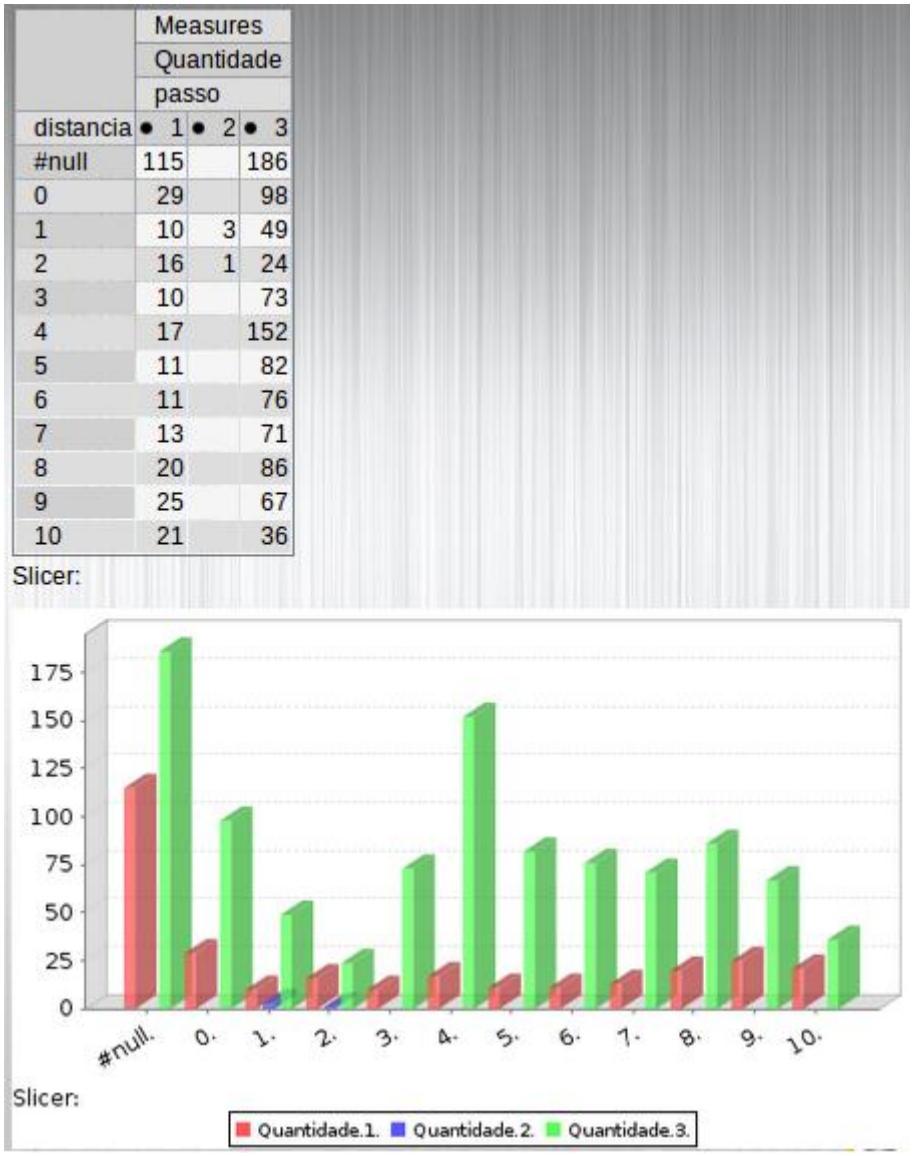
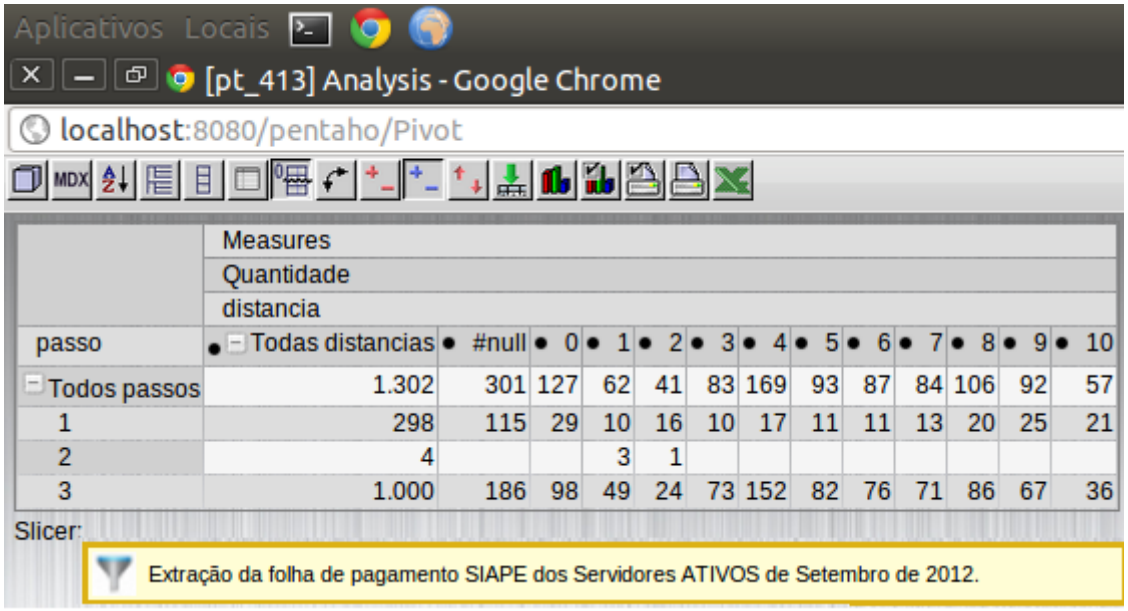


Figura 11 Qualidade de dados por meio de Levenshtein Distance

A Figura 12 traz a tabela analítica da qualificação dos dados obtidos por meio da distancia de Levenstein com os totalizadores, através dessa análise encontramos um total de 1302 registros corrigidos, o importante é observar que a limitação de 1000 registros para a terceira transformação limitou o potencial de correção desta intervenção.



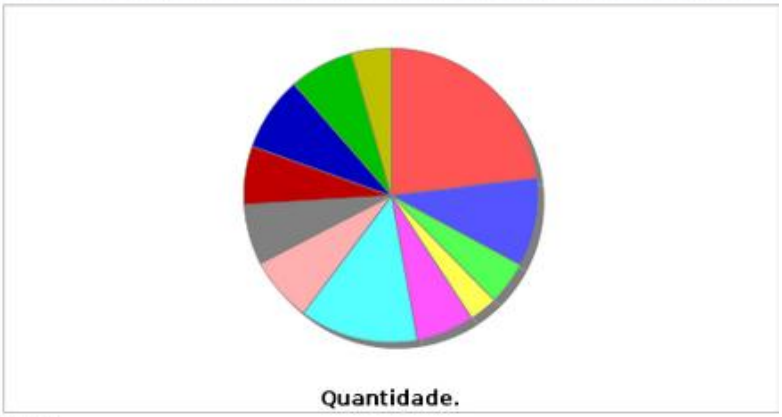
	Measures												
	Quantidade												
	distancia												
passo	Todas distancias	#null	0	1	2	3	4	5	6	7	8	9	10
Todos passos	1.302	301	127	62	41	83	169	93	87	84	106	92	57
1	298	115	29	10	16	10	17	11	11	13	20	25	21
2	4			3	1								
3	1.000	186	98	49	24	73	152	82	76	71	86	67	36

Figura 12 Analítico com totais por meio de *Levenshtein distance*

Para análise segue na Figura 13 o gráfico de pizza.

	Measures
distancia	Quantidade
#null	301
0	127
1	62
2	41
3	83
4	169
5	93
6	87
7	84
8	106
9	92
10	57

Slicer:



Slicer:

#null 0. 1. 2. 3. 4. 5. 6. 7. 8. 9. 10.

Figura 13 Gráfico pizza Levenshtein distance

4.2.2. Resultados obtidos por meio do Coeficiente de Dice

A Figura 14 traz a tabela analítica dos resultados obtidos por meio do Coeficiente de *Dice* e o gráfico de barra resultante. O gráfico traz as dimensões de coeficiente de 50% e de 85% de similaridade e ainda os resultados nos quais não foi possível encontrar similaridade com o resultado igual a *null*.

	Measures
	Quantidade
	passo
distancia	1	2	3
#null	115		186
0	29		98
1	10	3	49
2	16	1	24
3	10		73
4	17		152
5	11		82
6	11		76
7	13		71
8	20		86
9	25		67
10	21		36

Slicer:

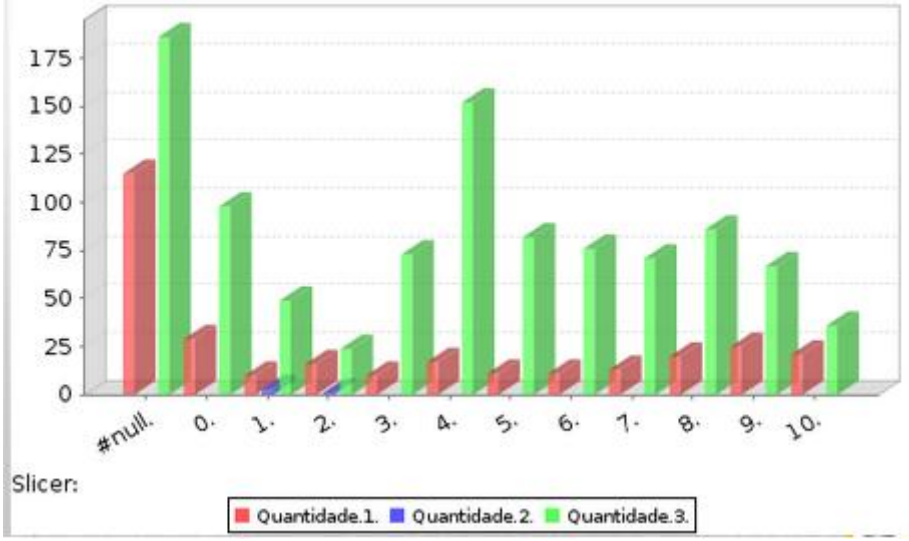
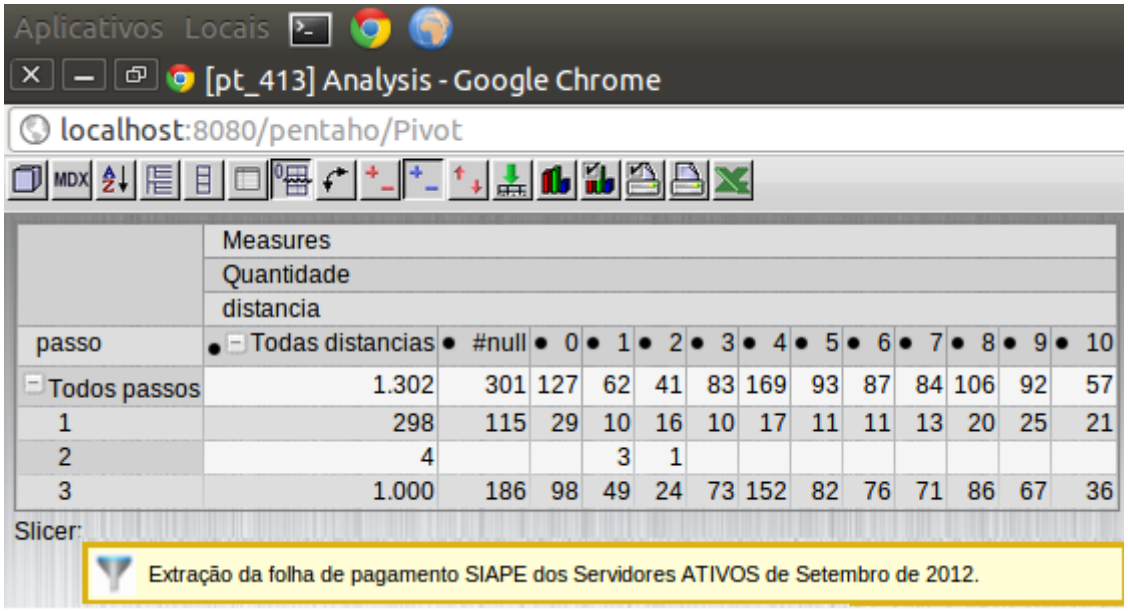


Figura 14 Qualidade de dados por meio do Coeficiente de Dice

A Figura 15 traz a tabela analítica da qualificação dos dados obtidos por meio da distancia de Levenstein com os totalizadores, através dessa análise encontramos um total de 1302 registros corrigidos, o importante é observar que a limitação de 1000 registros para a terceira transformação limitou o potencial de correção desta intervenção.



	Measures												
	Quantidade												
	distancia												
passo	Todas distancias	#null	0	1	2	3	4	5	6	7	8	9	10
Todos passos	1.302	301	127	62	41	83	169	93	87	84	106	92	57
1	298	115	29	10	16	10	17	11	11	13	20	25	21
2	4			3	1								
3	1.000	186	98	49	24	73	152	82	76	71	86	67	36

Slicer:

Extração da folha de pagamento SIAPE dos Servidores ATIVOS de Setembro de 2012.

Figura 15 Analítico com totais por meio do Coeficiente de Dice

Para análise segue na Figura 16 o gráfico de pizza.

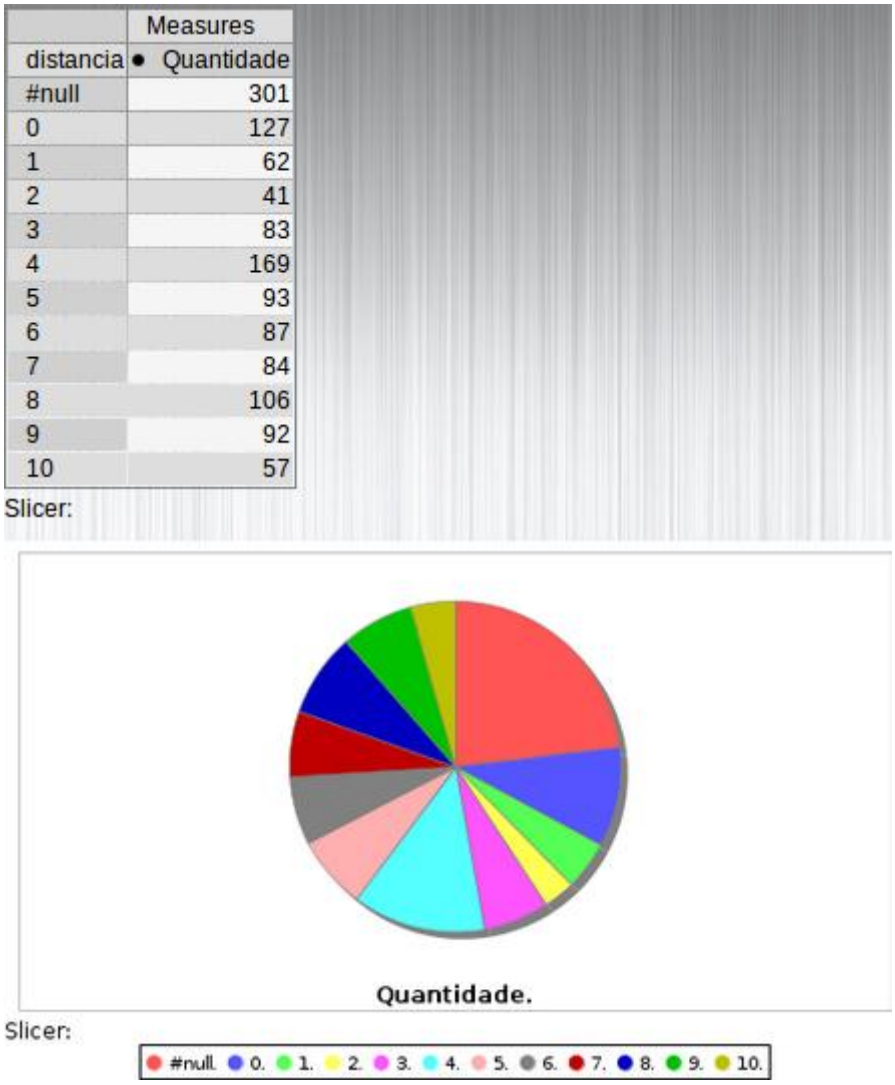


Figura 16 Gráfico Pizza Coeficiente de Dice

4.3. COMPARAÇÃO ENTRE OS RESULTADOS DAS TÉCNICAS APLICADAS

Além da implementação realizada em ambiente ETL, realizou-se uma aplicação na qual o interesse passou a ser a de fazer um comparativo entre os algoritmos de similaridade tratados nesta dissertação.

Esta implementação pode ser observada abaixo na Figura 17.

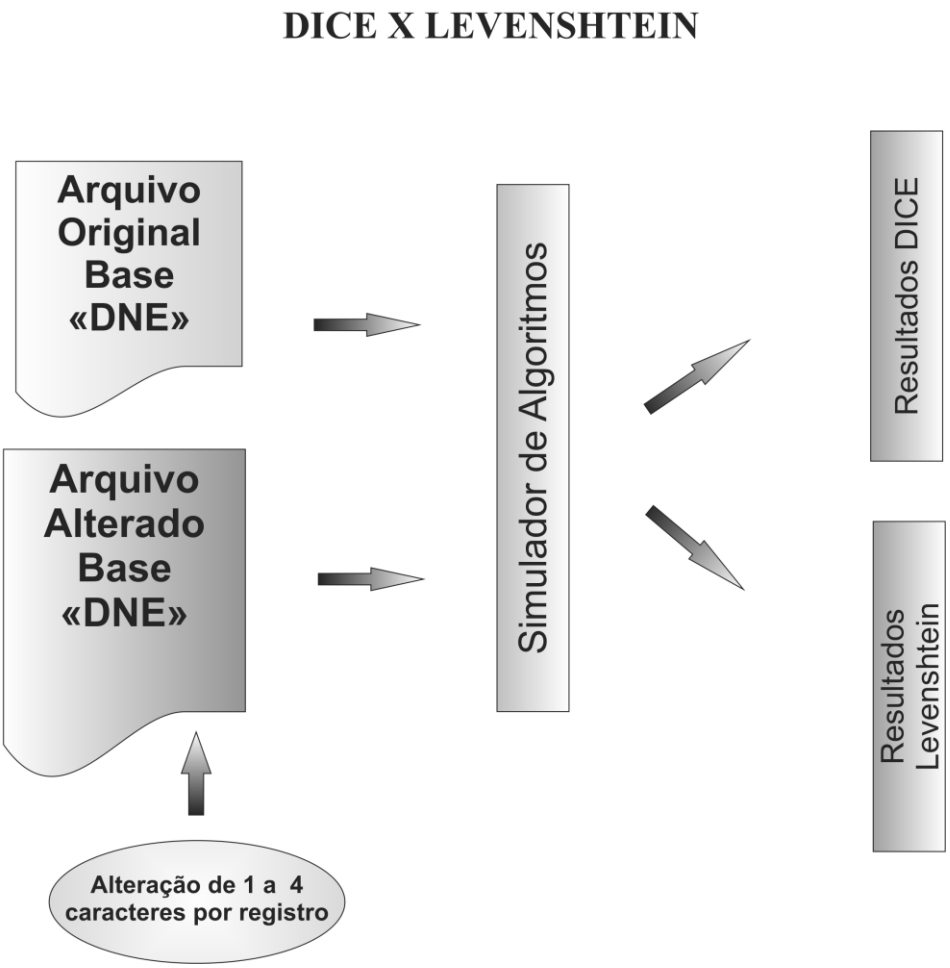


Figura 17 Comparação DICE X Levenshtein

Na Figura 17 foi ilustrado o processo que consolidou a comparação entre os algoritmos de Similaridade de Dice e de Distância de Levenshtein. Neste processo realizou-se três amostragens com número diferente de registros, na primeira retirou-se 100 registros de logradouros da base de dados do Diretório Nacional de Endereços (DNE) para efeito de validação da proposta, na segunda retirou-se 50 registros de logradouros da base de dados do Diretório Nacional de Endereços (DNE) para efeito de validação da proposta, comparação de tempo de resposta e acurácia do método, por fim retirou-se 20 registros de logradouros da base de dados do Diretório Nacional de Endereços (DNE) para efeito de validação da proposta, comparação de tempo de resposta e acurácia do método. Em um segundo passo estes arquivos contendo 100, 50 e 20 registros de logradouros foram submetidos a alteração manual em 50% de seus registros, nesta alteração o procedimento se deu aleatoriamente de maneira a alterar entre 1 a 4 caracteres do logradouro original representando 50% de alterações em cada número de amostragem, os resultados obtidos seguem na .

Tabela 2 Resultados obtidos Levenshtein X Dice

TÉCNICA APLICADA	LEVENSHTEIN		DICE	
TIPO DE RESULTADO / NÚMERO DE REGISTROS	ACURÁCIA (Percentual de corrigidos)	TEMPO DE RESPOSTA (seg)	ACURÁCIA (Percentual de corrigidos)	TEMPO DE RESPOSTA (seg)
20 REGISTROS	95%	0.1482 seg	70%	0.0911 seg
50 REGISTROS	98%	1,0495 seg	88%	0,7413 seg
100 REGISTROS	97%	3,7255 seg	76%	2,7671 seg

Os dois algoritmos foram devidamente codificados de maneira que pudessem ser testados em condições similares, dentro desta metodologia observaram-se os resultados devidamente elucidados como sendo acurácia o percentual de correções de logradouros, já como tempo de resposta considerou-se a quantidade de segundos que levou-se para concluir as ações.

No quesito referente a taxa de acurácia notou-se uma vantagem do algoritmo de distância de Levenshtein sobre o algoritmo de similaridade de Dice , sob todas as condições de números de registros submetidos, a saber 20, 50 e 100 registros, o resultado apresentado foi de 95%, 98% e 97%, respectivamente, para Levenshtein contra 70%, 88% e 76%, respectivamente, para Dice. No entanto faz-se necessário esclarecer que estes números se aplicam a esta experiência que teve um número relativamente baixo de alterações entre os arquivos comparados.

No que se refere ao tempo de resposta na execução dos algoritmos a vantagem se alterna, o desempenho, em termos de velocidade, pende para o algoritmo de similaridade de Dice, aplicado em todas as condições de amostragens, a saber 20, 50 e 100 registros, que apresentou os resultados de 0,0911 segundos, 0,7413 segundos e 2,7671 segundos, respectivamente, já o algoritmo de distância de Levenshtein apresentou os resultados de 0,1482 segundos, 1,0495 segundos e 3,7255 segundos, respectivamente. Nesta comparação nota-se que o tempo de resposta de Dice apresenta seus resultados em 61,47%, 70,63% e 74,27%, respectivamente do tempo decorrido na distância de Levenshtein. Como o algoritmo de Levenshtein é mais suscetível a número de interações necessárias para

realizar seu processamento, pode-se afirmar que a medida que a diferença entre o arquivo original e o arquivo a ser comparado, em termos de número de caracteres alterados, aumenta a distância de Levenshtein apresenta um desempenho, relativo a tempo de resposta, menos satisfatório em relação a similaridade de Dice.

5. MAPEAMENTO DE TRILHAS DE AUDITORIA UTILIZANDO INDEXAÇÃO ONTOLÓGICA

A proposta utiliza indicadores de auditoria como um instrumento de documentação das evidências, além de um meio para se controlarem os dados no momento em que são fornecidos e estruturados. Portanto, esta dissertação inclui uma metodologia padronizada de transformação de dados, por vezes não estruturados, da folha de pagamento em informações úteis para tomada de decisões relativas a irregularidades no âmbito do processo de auditoria da folha de pagamento.

Por meio de sua implementação como uma aplicação de *business intelligence*, a proposta é validada utilizando-se de dados originais gerados na análise da folha de pagamento mensal dos servidores do setor público brasileiro. Esta aplicação foi desenvolvida para auxiliar as atividades da Secretaria de Gestão Pública (SEGEP), pertencente ao Ministério do Planejamento, Orçamento e Gestão (MP).

Em sistemas de *Business Intelligence* (BI), os ambientes de *Data Warehouse* (DW) são capazes de armazenar enormes quantidades de dados extraídos de fontes diferentes. Este conjunto de dados compilados representam o conhecimento básico da organização e as decisões são tomadas com base em relatórios obtidos a partir destes dados.

O processo de ETL, quando realizado de maneira adequada, extrai os valores de distintos sistemas de origem e cumpre as normas de qualidade e consistência, para que os dados possam ser utilizados em conjunto [22]. Embora o processo de ETL seja transparente aos usuários finais, ele consome cerca de 70% dos recursos necessários para a implantação e manutenção de um típico DW, o que justifica o investimento em soluções mais eficientes para realização deste processo.

Para apoiar o processo de ETL, ontologias podem ser aplicadas para resolverem problemas de heterogeneidade de diferentes fontes de dados [46]. Em [47], o uso de ontologia combinado com mapas conceituais se mostra especialmente útil para facilitar o projeto conceitual dos bastidores de um DW. Em [48], é demonstrado que a aplicação de mapas conceituais pode melhorar o processo de concepção de um DW.

Nesta dissertação, propõe-se a aplicação de mapas conceituais no processo de ETL, a fim de detectar irregularidades em folhas de pagamento. Através da aplicação de mapas conceituais, as trilhas de auditoria podem ser validadas de acordo com as regras legais e o processo de ETL é utilizado para montar e apresentar os resultados; uma vez que os mapas conceituais são projetados de acordo com as regras legais, um processo de integração de dados correspondente pode ser utilizado no ETL para construção do DW.

Para validar a solução proposta, o arquivo fita espelho do banco de dados do SIAPE foi utilizado para o desenvolvimento do estudo de caso. Este sistema controla as folhas de pagamento de todos os funcionários públicos federais no Brasil.

O método original para este processo de auditoria na folha de pagamento baseava-se na análise e criação de planilhas ou pequenos bancos de dados que eram construídos manualmente pelos técnicos da área de auditoria da SEGEP/MP. Este método gerava um alto número de inconsistências e exigia muito tempo para validação dos dados. A fim de resolver estas limitações, na presente proposta de dissertação os mapas conceptuais são utilizados tanto para a validação das trilhas de auditoria como, posteriormente, para a execução do processo de ETL correspondente, incluindo-se sua carga automática dos dados em um DW.

O restante deste trabalho está organizado da seguinte forma. Na Seção 5.1, o processo de mapeamento de trilhas de auditoria é apresentado, enquanto na Seção 5.2, o processo de ETL é descrito. Os resultados obtidos na validação da solução proposta usando uma ferramenta de BI são apresentados na Secção 5.3.

5.1. PROCESSO ONTOLÓGICO DE FORMAÇÃO DE TRILHAS DE AUDITORIA

Para evitar ambiguidades e problemas de falta de padronização, a ontologia deve ser definida para cada domínio específico. Um sistema de aplicação de mapeamento ontológico, em que os conceitos são apresentados em caixas e as suas relações são identificados por um conjunto de frases conectadas, pode ser aplicado para esta tarefa.

Como estudo de caso para esta dissertação, a Seção 5.1.1 apresenta a trilha de auditoria estruturada considerando a incompatibilidade de rubricas de vencimento básico, enquanto a Seção 5.1.2 apresenta o mapeamento ontológico desta trilha de auditoria específica e os passos que a validam.

5.1.1. Trilha de auditoria: Incompatibilidade de vencimento básico

O controle da folha de pagamento dos servidores ativos, aposentados e pensionistas da Administração Pública Federal do Brasil baseia-se em leis, decretos e portarias.

O objetivo de cada trilha de auditoria é identificar alguma irregularidade no sistema SIAPE. Uma vez que é detectado um indício de irregularidade no pagamento, o sistema de apoio à auditoria deve notificar os auditores responsáveis para que eles possam tomar as ações necessárias no intuito de elucidar a referida inconsistência.

O elemento básico para o cálculo da folha de pagamento é chamado de rubrica. A rubrica é definida por um par de variáveis, sendo a primeira o nome e a segunda o valor da rubrica. Na folha de pagamento os valores das rubricas devem ser somados para compor o salário mensal pago a cada servidor público federal. Existem diversas rubricas, entre elas incluem-se: o vencimento básico, os impostos de receita, as contribuições para a seguridade social entre diversos outros. O sistema SIAPE possui um cadastro composto por 2.200 (duas mil e duzentas) rubricas, que podem ser de rendimento ou descontos, a serem aplicadas

mensalmente em um rol de aproximadamente 1.400.000 (um milhão e quatrocentos mil) servidores públicos federais.

Durante o desenvolvimento desta pesquisa, tendo como foco o seu objetivo e validação deste artigo, optou-se pela escolha da trilha de auditoria (01) "Controle de Incompatibilidade de Rubricas", que monitora atualmente 41 rubricas, dentre as quais, citamos a de Vencimento Básico (00001) que é incompatível com a percepção das rubricas de Provento Básico (00005), de Subsídio (82483) e Subsídio Aposentado (82484), conforme legislação [49] [50] [51].

5.1.2. Validação do mapeamento ontológico de uma trilha de auditoria específica: Incompatibilidade da rubrica Vencimento Básico

Buscou-se organizar todos os termos utilizando somente duas relações de inclusão: tem, é medido por. Obteve-se como resultado um mapa conceitual mais explícito, no qual se pode notar a hierarquia dos conceitos mais abrangentes em relação aos mais específicos, Este mapeamento ontológico é exemplificado na Figura 18 por um diagrama produzido com o software CmapTools [52] para integrar os conceitos de tempo, órgão, rubrica e o cargo do servidor.

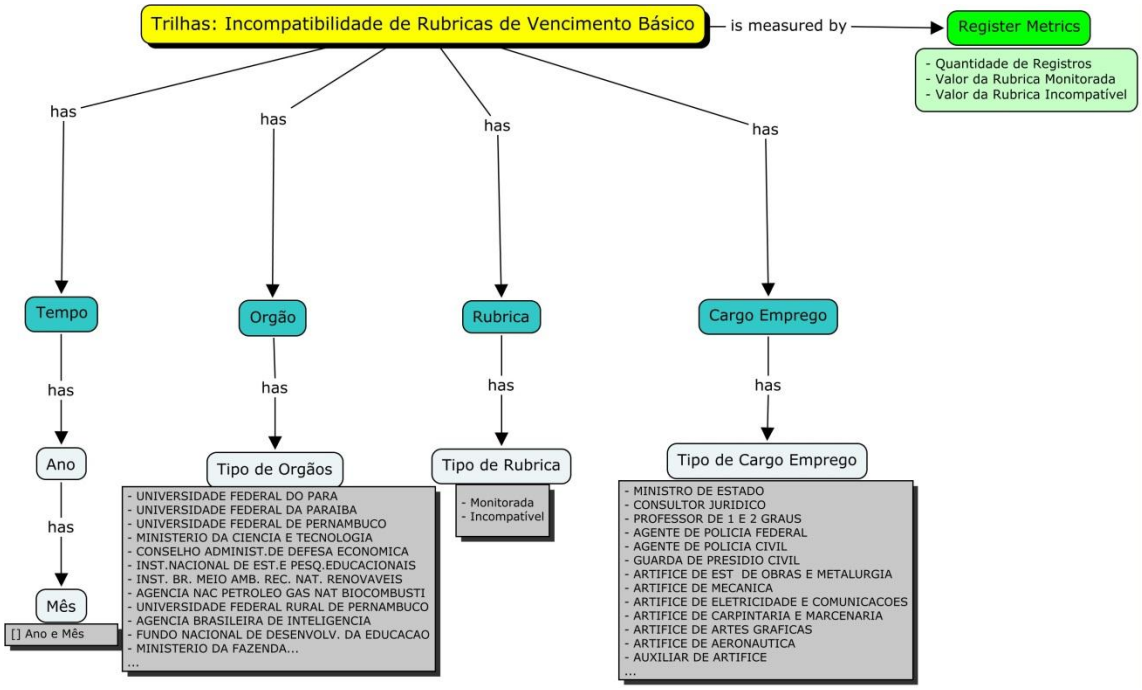


Figura 18 Mapeamento Ontológico

Esta figura corresponde a um modelo que possa seguir o raciocínio do auditor, auxiliando-o a correlacionar as dimensões de irregularidades da folha de pagamento, respondendo automaticamente a todos os questionamentos possíveis dentro das dimensões mapeadas aliadas às métricas cadastradas. Além disso, o mesmo diagrama pode ser implementado como um cubo multidimensional na aplicação de BI, de modo a transformar esta aplicação em uma ferramenta de auditoria automatizada fornecedora de informações bem definidas e úteis ao auditor.

Após o mapeamento ontológico, o processo segue os seguintes passos:

1. O processo inicia-se com a ação da Auditoria de Recursos Humanos (AUDIR) na extração dos dados do sistema SIAPE; estes dados incluem o Arquivo Espelho da base de dados do SIAPE e demais tabelas auxiliares que se integram na formação da folha de pagamento dos servidores públicos federais;

2. Realização do processo de ETL e descarga dos dados no Sistema Gerenciador de Banco de Dados (SGDB);

3. No próximo passo, o PDI realiza, de maneira periódica, o script que gera as trilhas de auditoria com os parâmetros de entrada baseados no ano e mês da folha que deve

ser verificado e especificamente qual trilha de auditoria está sendo focada. Para visualizar, de maneira mais clara, o script das trilhas de auditoria, criamos uma “raia” dentro deste mapeamento e, com isso, tratar especificamente da Trilha de Auditoria “Incompatibilidade de Rubricas” descrita neste artigo;

4. Inicia-se a busca dos servidores que possuam alguma movimentação financeira no ano e mês que a trilha está realizando sua geração;

5. Seguindo a descrição do script, verifica-se cada registro na busca por todos aqueles que possuem registro de exclusão, os quais são submetidos a outra condição que é encontrada por meio da busca pelos registros que atendam ao regime de situação “EST15” (instituidor de pensão) com o indicador de instituidor igual a zero; aqueles que não se enquadrarem nessa segunda condicional devem ser alijados do processo; retornando ao fluxo, tem-se o encontro dos registros que não possuem ocorrência de exclusão ou, se possuindo, encontram-se no Regime Jurídico “EST15” (com indicador de instituidor igual a zero); neste encontro se realiza a busca pelas rubricas incompatíveis;

6. Após essa busca, as inconsistências encontradas são gravadas em uma tabela específica; para os casos em que não são encontradas inconsistências para o respectivo ano e mês auditado, o processo é encerrado; nos casos onde as inconsistências foram encontradas e gravadas, o PDI gera e disponibiliza os dados dessas.

5.2. Implementação do processo de ETL

Para o processo de criação do DW, utilizam-se ferramentas de ETL que constituem grande parte dos fundamentos e o principal fragmento na construção de um DW.

O processo de ETL agrega valor significativo aos dados, pois se trata de um conjunto de ferramentas que realiza muito mais do que a agregação de dados pura e simples; além desta agregação, o sistema de ETL se destaca também por:

• Remover e corrigir erros de dados perdidos (dados fora do padrão);

• Fornecer medidas de confiança de maneira documentada e de maneira que possa ser rastreada;

• Realizar captura de dados transacionais do ambiente OLTP de maneira a serem disponibilizados em séries históricas;

• Ajustar dados de diversas fontes, padronizando-os para o ambiente de visualização (Ambiente OLAP);

• Estruturar e organizar valores para otimizar o ambiente OLAP;

A fim de lidar com grandes volumes de dados com qualidade, confiabilidade e capacidade de resposta em tempo hábil, já que se trata de uma base com aproximadamente dois milhões de registros, um *Service Level Agreement* (SLA) foi estabelecida de forma a responder às exigências da auditoria da folha de pagamento, executando o fluxo de informações sobre as trilhas de auditoria por meio de uma estrutura de processamento paralelo, como ilustrada na Figura 19. Esta estrutura permite a execução simultânea de diferentes componentes no fluxo de dados, abrindo caminho para melhorias no desempenho geral dos processos, bem como garantindo a qualidade e a confiabilidade das informações.

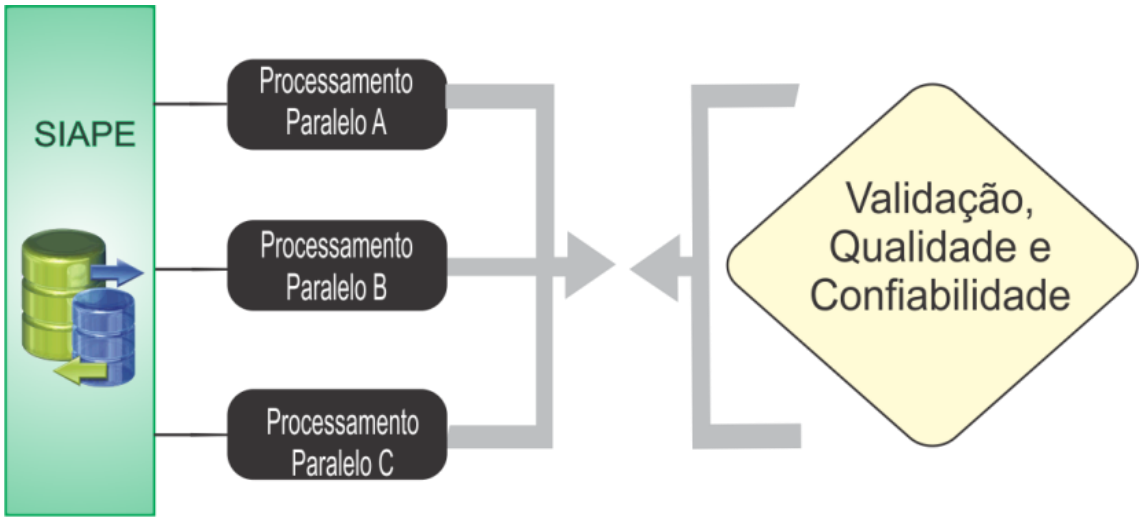


Figura 19 Processamento paralelo

5.2.1. Aplicação do *Operational Data Store* (ODS)

No processo de implementação de um DW, a construção de um ODS é facultativa, entretanto, ajuda em muito a diminuir os esforços de construção do DW. Todo o esforço de integração entre os sistemas transacionais da empresa seriam depositados no ODS e a carga do Warehouse seria consideravelmente simplificada.

Um ODS é integrado, orientado a assunto, volátil e estrutura tipo *current-valued* (valores atuais), desenhada para atender aos usuários operacionais, em grandes processos de integração, permitindo um melhor desempenho.

Na Figura 20, o ODS é visto como uma arquitetura que é alimentada por programas de transformação e integração (i/t). Estes programas de transformação e integração podem ser os mesmos programas que alimentam um DW ou programas separados. O ODS, por sua vez, alimenta um DW.

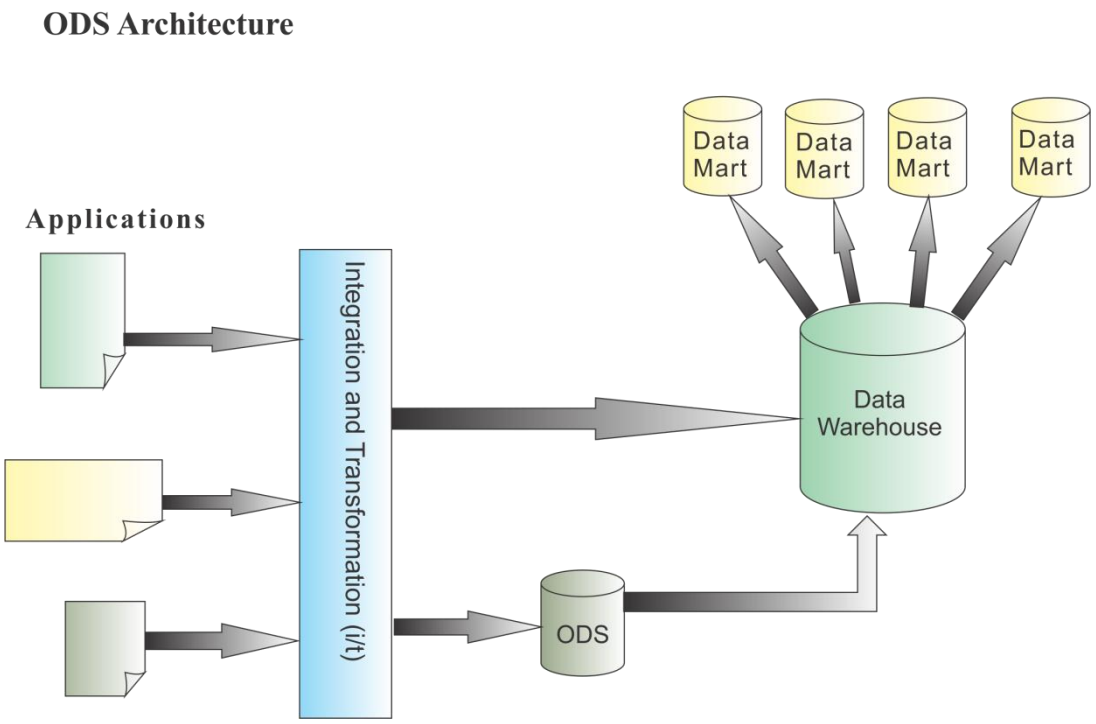


Figura 20 Arquitetura de um DW com Alimentação ODS

Para a construção dos filtros de indicadores que compõe o ODS das trilhas de auditoria utilizou-se *script* de consulta *Structured Query Language* (SQL), ou Linguagem de Consulta Estruturada. A Figura 21 apresenta, o modelo ODS criado para consultas nas trilhas de auditoria.

ods_trilhas	
♦id_ods_trilhas	integer
◦data_folha_arquivo_espelho	date
◦id_audir	varchar(4)
◦desc_trilha_auditoria	varchar(100)
◦cod_rubrica	varchar(5)
◦desc_rubrica	varchar(50)
◦cod_orgao	varchar(5)
◦desc_orgao	varchar(40)
◦upag	varchar(9)
◦mat_siape	varchar(7)
◦nome_servidor	varchar(100)
◦dat_entrada_ocupacao	date
◦car_emprego_classe	varchar(3)
◦ref_nivel_padrao	varchar(3)
◦sg_escolaridade	varchar(2)
◦cod_grupo_cargo	varchar(3)
◦cod_cargo_emprego	varchar(3)
◦desc_cargo_emprego	varchar(80)
◦des_situacao_funcional	varchar(50)
◦fun_sigla	varchar(3)
◦fun_cod_nivel	varchar(5)
◦id_ocorrencia_exclusao	integer
◦oco_exclusao_data	date
◦rejimejur	varchar(3)
◦codsituacaoservidor	varchar(2)
◦reg_obito_data	date
◦jor_trabalho	varchar(2)
◦carga_horaria_jornada	varchar(2)
◦cod_ocorrencia	text
◦oco_afasta_data_inicio	date
◦oco_afasta_data_termino	date
◦desc_ocorrencia	varchar(80)
◦num_propor	varchar(2)
◦denominador	varchar(2)
◦val_rubrica	numeric(12,2)
◦cod_rubrica_incompativel	varchar(5)
◦desc_rubrica_incompativel	varchar(50)
◦val_rubrica_incompativel	numeric(12,2)

Figura 21 Modelo ODS das trilhas de auditoria

Este processo finaliza-se com uma base de dados operacional, que passa a reagrupar as tabelas e informações, de modo a agregar as informações relevantes referentes ao

mapeamento das trilhas de auditoria; com isso poderá também gerenciar relacionamentos entre dados de diferentes origens com máxima granularidade com objetivo de criação do modelo dimensional.

5.2.2. Aplicação do *Data Warehouse (DW)*

Segundo [51], os componentes que formam um DW completo são:

• Sistemas de Origem – sistemas onde se encontram as fontes de dados, representam os locais de onde são extraídos todos os valores que irão se integrar a base de dados.

• *Data Staging Area* – essa é área que cria um ambiente intermediário de armazenamento e processamento de dados oriundos de diversas aplicações e fontes atuando de forma abrangente, desde o acesso à base dos dados nos sistemas de origem até a área de apresentação. Todas as Regras de negócio regem as transformações e organizações realizadas nesta área.

• Área de Apresentação de Dados – esta área responde pela interface onde o usuário realiza consultas, gera relatórios e outras aplicações gerenciais de análise sobre os dados devidamente organizados.

• Ferramenta de Acesso aos Dados – por meio de ferramentas de interação e visualização, os dados consolidados se tornam acessíveis e visíveis aos usuários.

O principal intuito na constituição de um DW é a capacidade de prover respostas completas e rápidas aos seus usuários, observando a máxima precisão dos resultados, a depender da qualidade da base de dados de origem, quanto ao tempo e inteligibilidade da resposta. Para isto, se faz necessário todo um planejamento baseado na necessidade de apresentação de resultados e informações.

A Figura 22 apresenta o modelo proposto para os indicadores da trilha de auditoria.

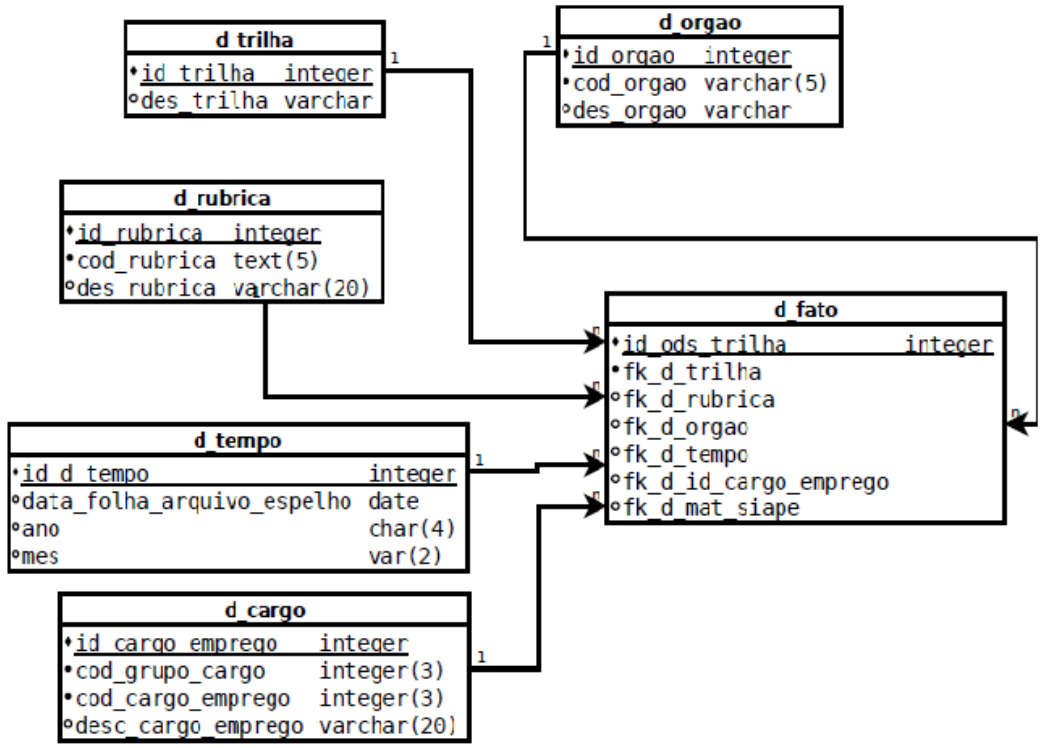


Figura 22 Modelo Floco de Neve

Observa-se uma dimensão complementar. No modelo floco de neve as tabelas dimensionais relacionam-se com a tabela de fatos, mas algumas dimensões relacionam-se apenas entre si, isto ocorre para fins de normalização das tabelas dimensionais, visando diminuir o espaço ocupado por elas. A implementação do floco de neve frustra o uso de esquemas de indexação mais eficientes como o índice de mapa de bits (55). Esses índices são muito úteis para indexar campos de baixa cardinalidade, agilizam muito o desempenho de uma consulta ou restrição à única coluna em questão, sendo assim ideais para tabelas sem normalização.

5.3. Validação utilizando a suíte *open source* Pentaho

A solução Pentaho define-se a si mesma como uma plataforma de BI orientada para a solução e centrada em processos. Ou seja, não só apresenta os resultados de uma forma

única e dando uma visão geral do estado da empresa, como programa os próprios processos (workflow) para a resolução de problemas detectados e apresentados.

Devido à sua estrutura em componentes, a Suíte pode ser utilizada para atender demandas que vão além do escopo das Soluções de BI mais tradicionais. Estão disponíveis componentes para a implantação de processos comandados por workflow automatizado, portais web customizáveis com suporte à *port lets* e *single sign-on*, entre outros.

Este objetivo guia a validação da proposta, conforme descrito a seguir, mas vale a pena ressaltar que as questões relacionadas com a integração inter-organizacional de domínio e distribuição de dados devem ser resolvidas a fim de compartilhar publicamente o significado dessas informações, bem como sua sintaxe e modos de utilização [53].

5.3.1. Apresentando os Resultados

Nas Figuras 53 e 54 apresenta-se a interface do Pentaho como ferramenta de inteligência e gestão com os resultados da trilha de auditoria objeto de estudo desta dissertação. Utilizaram-se três medidas representativas, sendo: quantidade de registros, valor da rubrica monitorada e o valor da rubrica incompatível.

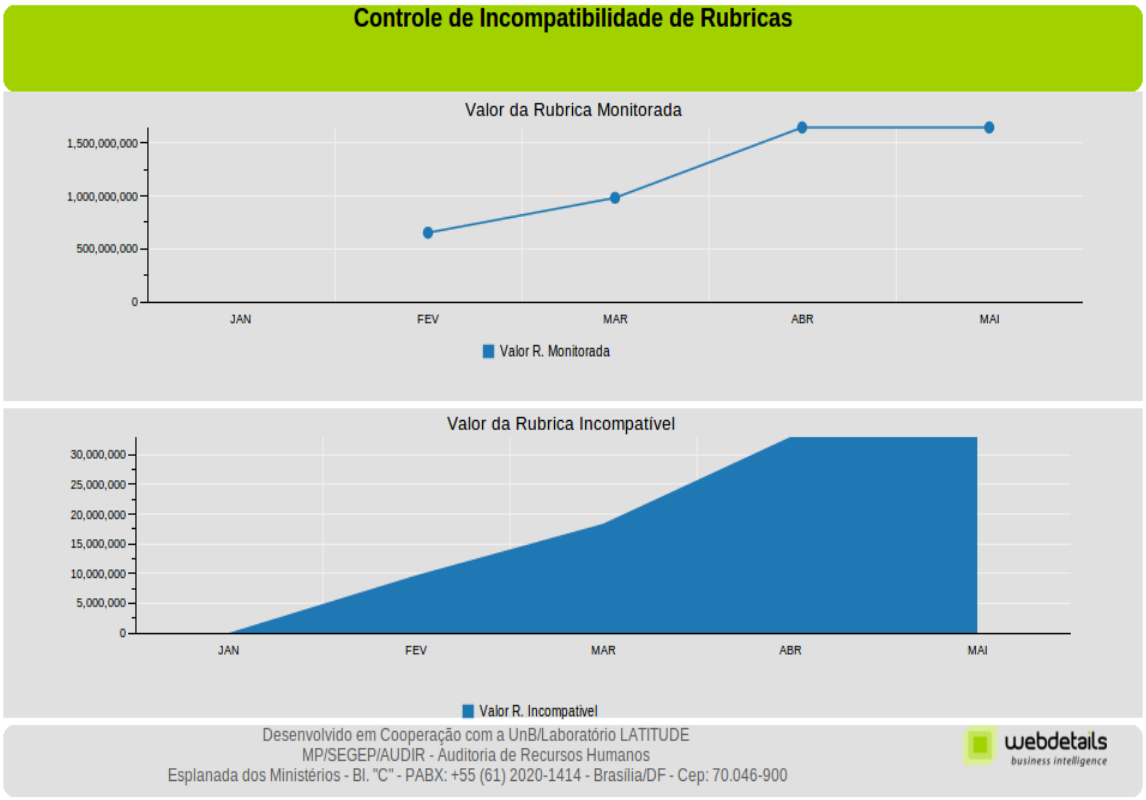


Figura 23 Dashboard Controle de Incompatibilidade de Rubrica

A Figura 23 e a Figura 24 fornecem o resultado do agrupamento e processamento de dados dispersos e não relacionados, que agora foram transformadas em conhecimento interessante, por meio da ontologia representada em mapas conceituais das trilhas de auditoria da folha de pagamento. Este processo leva em conta o objetivo estratégico de apoio a tomada de decisão da auditoria.

O gráfico da Figura 23 apresenta a evolução dos itens inconsistentes durante o período de fevereiro a junho de 2012. Esta mesma figura contém duas representações diferentes, um gráfico de linha e uma área de superfície, para as inconsistências detectadas pela ferramenta de BI. Estes gráficos indicam a eficácia no controle do sistema de folha de pagamentos durante este período.

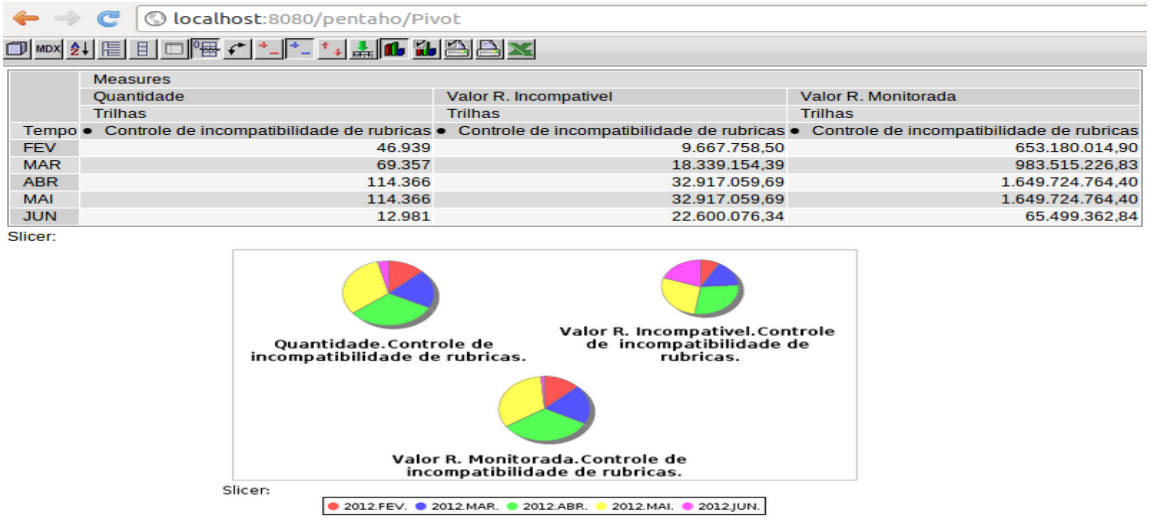


Figura 24 Gráfico do controle de incompatibilidade

Em alternativa, as tabelas na Figura 24 apresentam os resultados de uma representação analítica, especificando a quantidade de itens inconsistentes obtidos e os correspondentes valores que foram pagos durante o período, bem como o valor total dos itens monitorados durante este período. Ainda em relação à Figura 24, observam-se as representações gráficas de pizza correspondentes aos dados contidos na tabela da mesma figura. A vantagem desta visualização analítica é que ele oferece uma melhor interação para os gestores na execução de *drill up* e *drill down*, em suas tentativas de analisar as ocorrências com um valor maior ou menor e de acordo com a granulação e a temporalidade desejada.

Os resultados obtidos confirmam as melhorias para o processamento de trilhas de auditoria por meio da ontologia proposta combinada com mapas conceituais, o que facilita a compreensão do processo pelos auditores. Devido a esta validação, uma redução real das despesas na folha de pagamento dos servidores públicos federais do Brasil pode ser verificada, o que está diretamente relacionado a um melhor desempenho dos processos de auditoria da folha de pagamento.

6. CONCLUSÃO

No Ambiente da administração pública o contexto da utilização da qualificação de dados em ambientes de *Data Warehouses*, bem como criação de mapas conceituais que representam conceitos ontológicos se mostrou de extrema importância, representando a possibilidades de melhorias significativas no controle de gastos da folha de pagamento dos servidores públicos federais.

Durante a aplicação das técnicas de qualificação de dados em um ambiente de *Data Warehouse* verificou-se a utilização de duas técnicas.

A primeira, utilizando algoritmo da ferramenta de ETL automatizada. Apesar de sua utilização no processo de ETL, foi considerado a técnica com a utilização do algoritmo *levenshtein distance* embutido no próprio software da *Pentaho Data Integration.* Esta técnica apresentou resultados consistentes na busca pela qualificação dos dados disponibilizados no DW, foi verificado também que a aplicação desta técnica em ambientes com alta quantidade de registros apresenta dificuldades de desempenho, pois realiza grande interação em cada registro.

A segunda técnica utilizada de qualificação de dados utilizou-se de uma adaptação do Coeficiente de *Dice*, com passagem de parâmetros de 50% e 85% de similaridade. Este método se mostrou mais íntegro na geração de resultados mais confiáveis. Como esta técnica foi desenvolvida e configurada integralmente por meio da linguagem de programação Java, foi possível otimiza-la de forma a atender aspectos de desempenho e utilização de hardware, o que proporcionou um melhor desempenho e confiança nos dados obtidos. Depois de desenvolvida a codificação, esta foi inserida durante o processo de ETL dentro da arquitetura de sua ferramenta.

O trabalho também apresentou-se como método eficaz a aplicação do mapeamento ontológico das Trilhas de Auditoria. Foi essencial a verificação da informação desde os seus dados de origem até a apresentação de resultados para os usuários em formato de

relatórios gerenciais, por meio de um processo de geração de Trilhas de Auditoria com a utilização de ontologia construída por meio da técnica de mapas conceituais.

Com a fundamentação destas informações possibilitou-se a prévia visualização das informações a serem emitidas na folha de pagamento, de modo a atuar de maneira preventiva na correção de indícios de irregularidades permitindo que esta intervenção seja realizada antes da confecção da folha de pagamento, contribuindo assim com a diminuição das incertezas na auditoria das contas públicas. Esta técnica já se encontra em aplicação no Ministério do Planejamento, apoiando o processo de auditoria na folha de pagamento dos servidores públicos federais do Brasil.

Tendo em vista os números apresentados pelo SIAPE:

- *1,4 milhões de servidores públicos federais (ativos, aposentados e pensionistas);*
- *Sistema com mais de dois milhões de registros;*
- *214 órgãos;*
- *240 campos;*
- *Folha Mensal girando em torno de 15 Bilhões de Reais;*
- *Apontamentos de suspeitas de irregularidades nas rubricas incompatíveis monitoradas da ordem de 30 milhões de reais mensais;*
- *Valor do Projeto de Pesquisa junto ao CDT/UnB em torno de um milhão de reais ao ano;*

Verificou-se o potencial de ganho pecuniário junto aos cofres públicos, o que por si só justificaria o projeto de pesquisa aplicada.

6.1. SUGESTÕES PARA TRABALHOS FUTUROS

Como proposta de trabalhos futuros, podem ser consideradas as seguintes evoluções:

- Pesquisa e análise sobre os demais algoritmos de qualificação de dados sobre os demais dados do arquivo espelho do SIAPE;
- Análise e testes com outros métodos de similaridades de *strings*, na busca por um método mais eficaz, sem deixar de testar cada parâmetro que alimenta estes métodos;
- O incremento da ontologia sobre as demais trilhas de auditoria utilizando as técnicas de mapas conceituais como proposto neste trabalho, com o objetivo de homogeneizar a compreensão de outras áreas da gestão de negócios e modelagem de dados;
- Pesquisa e aplicação de algoritmos de predição sobre possíveis falhas de carga na alimentação de sistemas.
- Notou-se grande aplicabilidade dos algoritmos de índice de similaridade entre *strings* no processo de qualificação de DW. Entretanto como nesta dissertação as técnicas foram apresentadas de maneira independentes. A proposta é um estudo mais amplo no qual uma técnica de algoritmo complemente a outra. A saída qualificada de um algoritmo possa ser a entrada de dados do seguinte, incluindo-se os algoritmos de Dice, Levenshtein e demais que possam se encaixar nesta proposta.

7. REFERÊNCIAS BIBLIOGRÁFICAS

[1] S. R. CAMPOS, A. A. FERNANDES, R. T. SOUSA JUNIOR, E. P. FREITAS, J.P.C.L. da Costa e A. M. SERRANO, "ONTOLOGIC AUDIT TRAILS MAPPING FOR DETECTION OF IRREGULARITIES IN PAYROLLS," em *Conference Record, industry Applications Society, IEEE-IAS Anual Meeting*, São Carlos, Brasil, 2012.

[2] W. J. OLIVEIRA, Data Warehouse, 2 ed., Florianópolis: Visual Books, 2002.

[3] F. V. PRIMAK, Decisões com B.I., São Paulo: Ciencia Moderna, 2008.

[4] J. C. KENYON, G. S. AEBY, R. E. BRAINARD, J. D. CHOJNACKI, M. DUNLAP e C. B. WILKINSON, "MASS CORAL BLEACHING ON HIGH LATITUDE REEFS IN THE HAWAIIAN ARCHIPELAGO," em *10th International Coral reef Symposium*, Okinawa, Japan, 2004.

[5] T. S. BATEMAN e S. A. SNELL, MANAGEMENT: COPETING IN THE NEW ERA, 5 ed., Garland, U.S.A.: Irwin Professional PUB, 2002.

[6] C. CALDO, "HOMOLOGAÇÃO DE MODELOS DE DADOS - AVALIANDO A QUALIDADE NA RAIZ DO PROJETO," *SQL MAGAZINE*, n. 22, pp. 62-66, 2005.

[7] M. SCANNAPIECO e T. CATARCI, "DATA QUALITY UNDER THE COMPUTER SCIENCE PERSPECTIVE," em *DIPARTIMENTO DI INFORMATICA E SISTEMISTICA,UNIVERSITA DI ROME*, Roma, Itália, 2002.

[8] G. BRACKSTONE, "HOW IMPORTANT IS AACCURACY?," em *PROCEEDINGS OF STATISTICS SYMPOSIUM*, Canada, 2001.

[9] L. ENGLISH, "TEN YEARS OF INFORMATION QUALITY ADVANCES: WHAT NEXT?," em *INFORMATION IMPACT INTERNATIONAL, INC*, 2001.

[10] M. D. Hansen, "Zero defect data," em *Massachusetts Institute of Technology*, Massachusetts, 1991.

[11] "http://mitiq.mit.edu," [Online].

[12] C. T. Redman, Data Quality Management and Technology, New York: Bantam Books, 1994.

[13] W. ECKERSON, DATA QUALITY AND THE BOTTOM LINE: ACHIEVING BUSINESS SUCCESS THROUGH A COMMITMENT TO HIGH QUALITY DATA, Seattle, USA: The Data Warehousing Institute, 2002.

[14] R. WANG, Y. LEE e B. DAVIDSON, "DEVELOPING DATA PRODUCTION MAPS: MEETING PATIENT DISCHARGE DATA SUBMISSION REQUIREMENTS," em *HEALTCARE TECHNOLOGY AND MANAGEMENT*, 2004.

[15] J. KYL, "THE DATA QUALITY ACT: HISTORY AND PURPOSE," em *UNITED STATES SENADE - REPUBLICAN POLICY COMMITTEE*, USA, 2005.

[16] W. H. INMON, COMO CONSTRUIR O DATA WAREHOUSE, 2 ed., Rio de Janeiro: CAMPUS, 1997.

[17] R. KIMBALL, L. REEVES, M. ROOS e W. THORNTHWAITE, THE DATA WAREHOUSE LIFECYCLE TOOLKIT, New York, USA: JOHN WILEY & SONS, 1998.

[18] V. POE, P. KLAUER e S. BROBST, BUILDING A DATA WAREHOUSE FOR DECISION SUPPORT, PRENTICE HALL, 1998.

[19] X. WU e D. BARBARÁ, "MISSING VALUES FROM SUMMARY CONSTRAINTS," *ACM SIGKDD EXPLORATION NEWSLETTER*, vol. 4, 2002.

[20] P. BUNEMAN, S. KHANNA e W. TAN, "WHY AND WHERE: A CHARACTERIZATION OF DATA PROVENANCE," em *INTERNATIONAL CONFERENCE ON DATABASE THEORY*, London, England, 2001.

[21] R. KIMBALL e W. ROSS, THE DATA WAREHOUSE TOOLKIT: THE COMPLETE GUIDE TO DIMENSIONAL MODELING, JOHN WILEY & SONS, INC, 2002.

[22] R. KIMBALL e J. CASSERTA, THE DATA WAREHOUSE ETL TOOLKIT: PRACTICAL TECHNIQUES FOR EXTRACTING, CLEANING, CONFORMING AND DELIVERING DATA, Indianapolis, USA: WILEY PUBLISHING, 2004.

[23] C. BARBIERI, BI - BUSINESS INTELLIGENCE: MODELAGEM E TECNOLOGIA, Rio de Janeiro: AXCEL BOOKS, 2001.

[24] A. SERRANO e C. FILHO, GESTÃO DO CONHECIMENTO, FCA - EDITORA DE INFORMÁTICA, 2003.

[25] Y. WAND e R. WANG, "ANCHORING DATA QUALITY DIMENSIONS IN ONTOLOGICAL FOUNDATIONS," *COMMUNICATIONS OF THE ACM*, vol. 39, n. 11, pp. 86-95, 1996.

[26] L. AMARAL e J. VARAJÃO, PLANEJAMENTO DE SISTEMAS DE INFORMAÇÃO, FCA - EDITORA DE INFORMÁTICA, 2000.

[27] M. BOUZEGHOUB e V. PERALTA, "A FRAMEWORK FOR ANALYSIS OF DATA FRESHNESS," *COMUNICATIONS OF THE ACM*, pp. 59-67, 2004.

[28] INMON, BUILDING THE DATA WAREHOUSE, New York, USA: JOHN WILEY & SONS, INC, 1996.

[29] G. AMARAL, "AQUAWARE: UM AMBIENTE DE SUPORTE À QUALIDADE DE DADOS EM DATA WAREHOUSE," em *DISSERTAÇÃO DE MESTRADO, UFRJ*, Rio de Janeiro, 2003.

[30] J. RASCÃO, SISTEMAS DE INFORMAÇÃO PARA ORGANIZAÇÕES, Lisboa, Portugal: SÍLABO, 2001.

[31] K. RASMUSSEM, "ELEMENTARY DATA QUALITY ELEMENTS," em *IASSIST 2004*, Marison, USA, 2004.

[32] R. WANG, D. STRONG e L. GUARASCIO, "BEYOND ACCURACY: WHAT DATA QUALITY

MEANS TO DATA CONSUMERS," em *TDQM RESEARCH PROGRAM, SLOAN SCHOOL OF MANAGEMENT, MASSACHUSETTS INSTITUTE OF TECHNOLOGY.*, Cambridge, USA, 1994.

[33] Y. LEE, M. JARKE, S. MADNICK, Y. WAND, J. FUNK e P. BOWEN, "DATA QUALITY IN INTERNET TIME, SPACE AND COMMUNITIES," em *ICIS 2000 PROCEEDINGS OF THE TWENTY INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS*, Atlanta, USA, 2000.

[34] A. A. FERNANDES, L. C. AMARO, J. C. da COSTA, A. R. SERRANO, V. A. MARTINS e R. T. de SOUSA Jr, "CONSTRUCTION OF ONTOLOGIES BY USING CONCEPT MAPS: A STUDY CASE OF BUSINESS INTELLIGENCE FOR THE FEDERAL PROPERTY DEPARTMENT," em *INTERNATIONAL CONFERENCE ON BUSINESS INTELLIGENCE AND FINANCIAL ENGINEERING (BIFE'12)*, Lanzhou & Tunhuang, China, 2012.

[35] M. HELFERT e C. HERRMANN, "PROACTIVE DATA QUALITY MANAGEMENT FOR DATA WAREHOUSESYSTEMS - A METADATA BASAED DATA QUALITY SYSTEM," em *DATA WAREHOUSING 2, INSTITUTE OF INFORMATION MANAGEMENT, UNIVERSITY OF ST. GALLEN* , 2002.

[36] K. ORR, "DATA QUALITY AND SYSTEMS THEORY," *COMMUNICATIONS OF THE ACM*, vol. 41, n. 2, pp. 66-71, 1998.

[37] L. ENGLISH, IMPROVING DATA WAREHOUSING AND BUSINESS INFORMATION QUALITY, New York, USA: JOHN WILEY & SONS, INC, 1999.

[38] Z. BAR-YOSSEF, "APPROXIMATING EDIT DISTANCE EFFICIENTLY," em *PROCEEDINGS. 45TH ANNUAL IEEE SYMPOSIUM*, Haifa, Israel , 2004.

[39] T. V. OLIVEIRA, "ADAPTAÇÃO DO ALGORITMO LEVENSHTEIN DISTANCE," em *UNIVERSIDADE CATÓLICA DE PELOTAS*, Pelotas, 2009.

[40] L. R. DICE, MEASURES OF THE AMOUNT OF ECOLOGIC ASSOCIATION BETWEEN SPECIES, 1945.

[41] G. KONDRAK, D. MARCU e K. KNIGHT, "COGNATES CAN IMPROVE STATISTICAL TRANSLATION MODELS," em *PROCEEDING OF HLT-NAACL*, 2003.

[42] E. RAHM e H. H. DO, "DATA CLEANING: PROBLEMS AND CURRENT APPROACHES," *IEEE BULLETIN OF THE TECHNICAL COMMITTEE ON DATA ENGINEERING*, vol. 23, n. 4, 2000.

[43] M. GONÇALVES, EXTRAÇÃO DE DADOS PARA DATA WAREHOUSE, Rio de Janeiro: AXCEL BOOKS, 2003.

[44] T. GRUBER, "A TRANSLATION APPROACHES TO PORTABLE ONTOLOGY SPECIFICATIONS," *KNOWLEDGE ACQUISITION*, vol. 5, n. 2, pp. 199-220, 1993.

[45] M. AMORETTI e L. TAROUCO, "MAPAS CONCEITUAIS: MODELAGEM COLABORATIVA DO CONHECIMENTO," Porto Alegre, Brasil, 2000.

[46] Y. DING e S. FOO, "ONTOLOGY RESEARCH AND DEVELOPMENT: PART 1 – A REVIEW OF," *JOURNAL OF INFORMATION SCIENCE*, 2002.

[47] N. GUARINO, "UNDERSTANDING, BUILDING AND USING ONTOLOGIES," *JOURNAL HUMAN-COMPUTER STUDIES*, 1997.

[48] N. GUARINO, "FORMAL ONTOLOGIES AND INFORMATION SYSTEMS," em *INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY IN INFORMATION SYSTEMS FOIS'98*, Trento, Italy, 1998.

[49] BRASIL, *VENCIMENTO BÁSICO - LEI N º. 8.112*, 1990.

[50] BRASIL, *PROVENTO BÁSICO - LEI N º. 8.112*, 1990.

[51] BRASIL, *SUBSÍDIO - CÓDIGOS: 82484, 82483, 82286, LEI N º. 11.358*, 2006.

[52] "CMAP TOOLS DOCUMENTATIONS – IHMC," [Online]. Available: http://cmap.ihmc.us/documentation.

[53] V. A. MARTINS, J.P.C.L. da COSTA e R. T. de SOUSA Jr, "Architecture of a Collaborative Business Intelligence Environment based on an Ontology Repository and Distributed Data Services," em *International Conference on Knowledge Management and Information Sharing (KMIS 2012)*, Barcelona, SPAIN, 2012.

[54] J. T. ROSS, FUZZY LOGIC WITH ENGINEERING APPLICATIONS, 3 ed., JOHN WILEY & SONS, 2010.

[55] L. A. ZADEH, "IS THERE A NEED FOR FUZZY LOGIC?," *INFORMATION SCIENCES*, vol. 178,

pp. 2715-2779, 2008.

[56] L. A. ZADEH, "FUZZY SETS," *INFORMATION AND CONTROL,* vol. 8, pp. 338-353, 1965.

[57] J. VARAJÃO, ARQUITECTURA DA GESTÃO DE SISTEMAS DE INFORMAÇÃO, FCA - EDITORA DE INFORMÁTICA, 1998.

[58] T. REDMAN, "DATA: AN UNFOLDING QUALITY DISASTER," *DM REVIEW MAGAZINE*, 2004.

[59] "www.egi.ua.pt," [Online].

[60] H. WATSON, B. WIXOM, D. ANNINO, K. AVERY e M. RUTHERFORD, "CURRENT PRACTICESS IN DDATA WAREHOUSING," *DATA WAREHOUSE TODAY*, 2001.

[61] M. BRACKETT, THE DATA WAREHOUSE CHALLENGE, New York, USA: JOHN WILEY & SONS, INC, 1996.

[62] M. HELFERT e E. MAUR, "A STRATEGY FOR MANAGINGDATA QUALITY IN DATA WAREHOUSE SYSTEMS," *INSTITUTE OF MANAGEMENT, UNIVERSITY OF ST. GALLEN*, 2011.

[63] W. KIM, B. CHOI, E. HONG, S. KIM e D. LEE, "A TAXONOMY OF DIRTY DATA," *DATA MINING AND KNOWLEDGE DISCOVERY*, n. 7, pp. 81-99, 2003.

[64] P. OLIVEIRA, F. RODRIGUES, P. HENRIQUES e H. GALHARDAS, "A TAXONOMY OF DATA QUALITY PROBLEMS," em *PROCEEDINGS OF 2nd INTERNATIONAL WORKSHOP ON DATA QUALITY*, Porto, Portugal, 2005.

[65] P. VASSILIADIS, M. BOUZEGHOUB e C. QUIX, "TOWARDS QUALITY-ORIENTED DATA WAREHOUSE USAGE AND EVOLUTION," em *PROCEEDINGS 11th CONFERENCE OF ADVANCED INFORMATION SYSTEMS ENGINEERING (CAISE ' 99)*, Heidelberg, Germany, 1999.

[66] M. JARKE, U. GRIMMER e L. DOMINIK, "SYSTEMATIC DEVELOPMENT OF DATA MINING - BASED DATA QUALITY TOOLS," em *PROCEEDINGS OF THE 29th VERY LARGE DATABASES CONFERENCE*, Berlim, Germany, 2003.

[67] R. WANG, Y. LEE, L. PIPINIO e D. STRONG, "MANAGE YOUR INFORMATION AS A PRODUCT," *SLOAN MANAGEMENT REVIEW*, vol. 39, pp. 95-105, 1998.

[68] L. PIPINO, Y. LEE e R. WANG, "DATA QUALITY ASSESSMENT," *COMMUNICATIONS OF THE ACM,* vol. 45, n. 4, pp. 211-218, 2002.

[69] R. WANG, "A PRODUCT PERSPECTIVE ON TOTALQUALITY MANAGEMENT," *COMMUNICATIONS OF DE ACM,* vol. 41, n. 2, pp. 58-65, 1998.

[70] D. STRONG, Y. LEE e R. WANG, "DATA QUALITY EN CONTEXT," *COMMUNICATIONS OF THE ACM,* vol. 40, n. 5, pp. 103-110, 1997.

[71] A. J. FARIA e J. R. DICKINSON, "THE MARKETING MANAGEMENT SIMULATION," em *LASALLE, ON: THE SIMULATION SOURCE*, 1995.

[72] J. D. NOVAK, "THE THEORY UNDER LAYING CONCEPT MAPS AND HOW TO CONSTRUCT THEM," 2003.

APÊNDICE A - PUBLICAÇÕES REALIZADAS DURANTE O MESTRADO

1.	**CAMPOS, S. R.** ; FERNANDES, A. A. ; SOUSA JUNIOR, R. T. ; FREITAS, E. P. ; J. P. C. L. da Costa ; SERRANO, A. M. R. . ONTOLOGIC AUDIT TRAILS MAPPING FOR DETECTION OF IRREGULARITIES IN PAYROLLS. Conference Record, Industry Applications Society, IEEE-IAS Annual Meeting, 2012.

ANEXO A

ANEXO A – APLICAÇÃO FUZZY MATCH COM ALGORITMO DE LEVENSHTEIN

Apresenta-se abaixo o detalhamento dos processos de ETL implementados para realização do estudo de caso, utilizando-se da aplicação "*Fuzzy Match*" configurado para aplicação em algoritmo de Levenshtein:

A Figura 25 – Transformação 1 - Encontrar CEP por meio de algoritmo de Levenshtein realizada por meio da ferramenta de ETL – PDI.

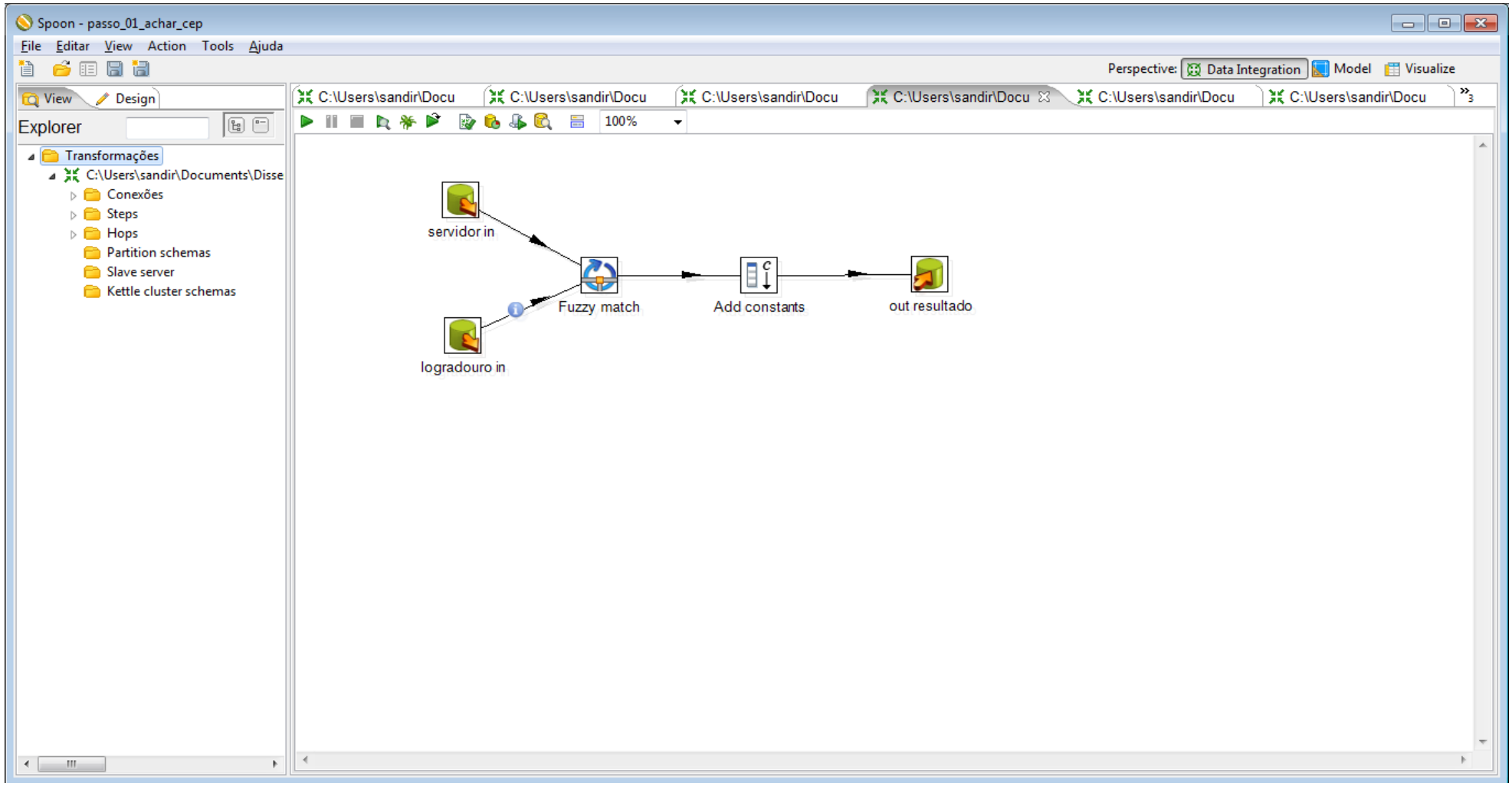


Figura 25 – Transformação 1 - Encontrar CEP por meio de algoritmo de Levenshtein

A transformação inicia-se com no passo da Figura 26 Table Input servidor in.

Figura 26 Table Input servidor in

Na ferramenta de ETL – PDI o passo do tipo "*Table Input*" é utilizado para realizar a leitura das informações das fontes de dados que alimentam a transformação.

Ao se abrir o passo servidor in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados gerado pela extração do arquivo espelho do SIAPE na tabela siape_servidor como visto na Figura 27 Servidor in.

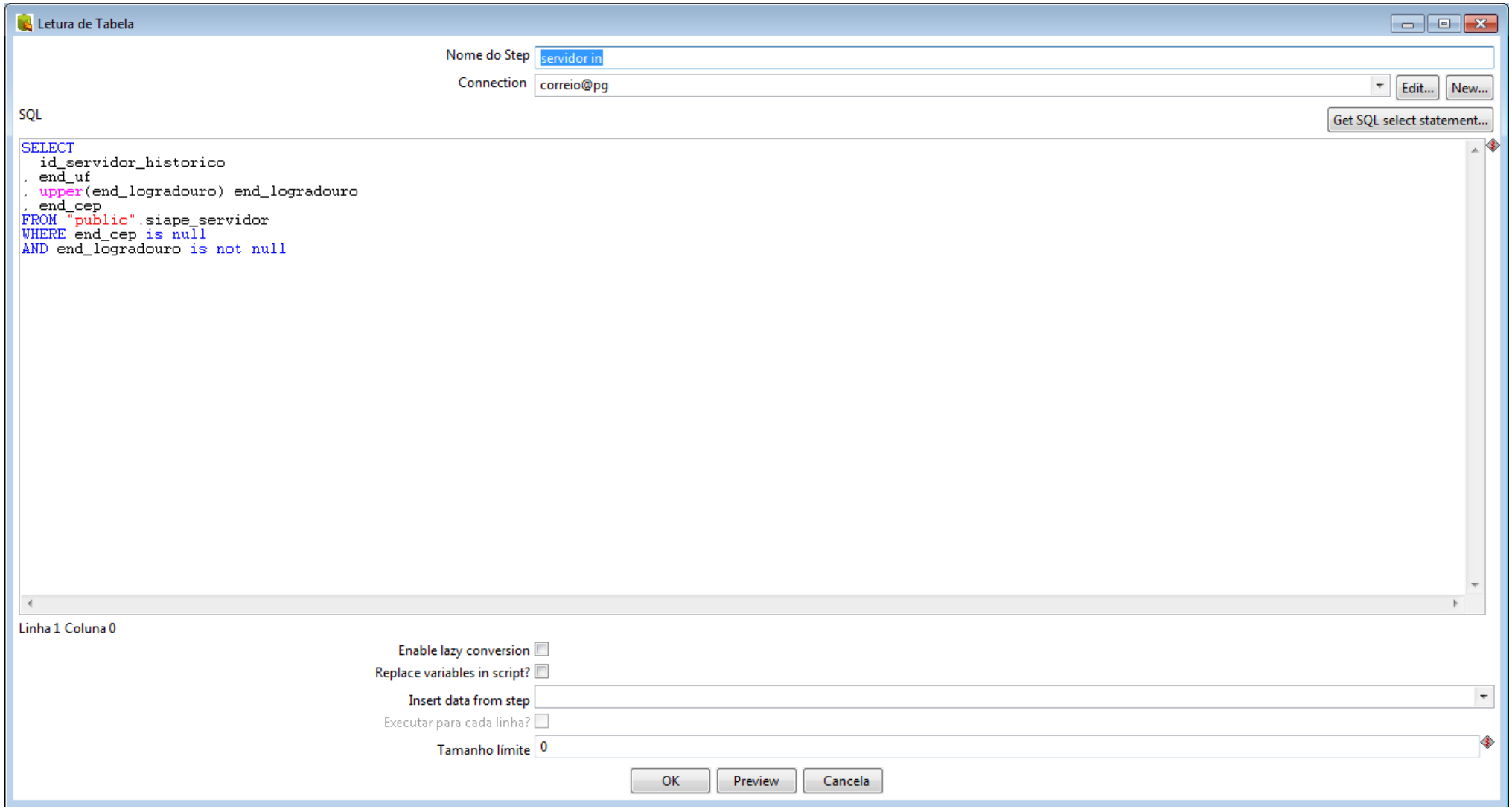


Figura 27 Servidor in

Como visto na Figura 27, observa-se que para a realização da “Transformação 1 Encontrar CEP por meio de algoritmo de Levenshtein”, o comando SQL executado na tabela siape_servidor traz os seguintes campos: id_servidor_historico (identificador único do servidor), end_uf (UF do endereço do servidor), end_logradouro em caixa alta para que possa ser comparada com a base do DNE e por fim o campo end_cep (traz o CEP cadastrado para o servidor).

Verificou-se a necessidade de utilização do comando “upper” no intuito de transformar todos os caracteres de um campo do tipo “string” em caracteres de caixa alta, na Figura 28 exibe-se uma consulta realizada sem o comando “upper” onde pode ser verificado em destaque um endereço de logradouro em caixa baixa.

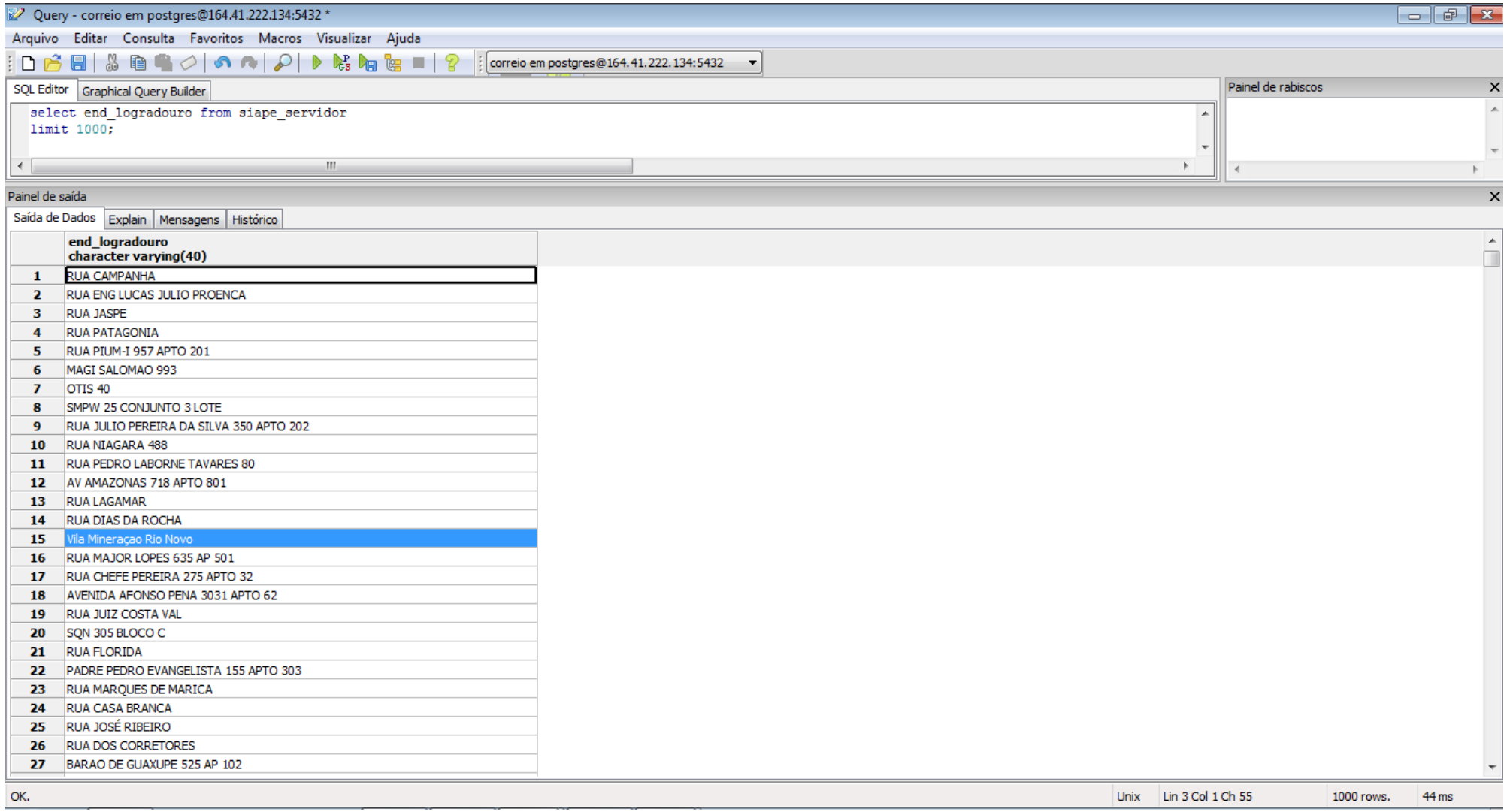


Figura 28 Demonstração da utilização do comando upper

O comando SQL de seleção realizado, também condiciona aos registros que possuírem o campo end_cep sem nenhum valor inserido e o campo end_logradouro com algum valor inserido.

A transformação continua com o passo do tipo *Table Input* logradouro in Figura 29.

Figura 29 Table Input logradouro in

Ao se abrir o passo logradouro in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados do DNE dentro da tabela log_logradouro Figura 30.

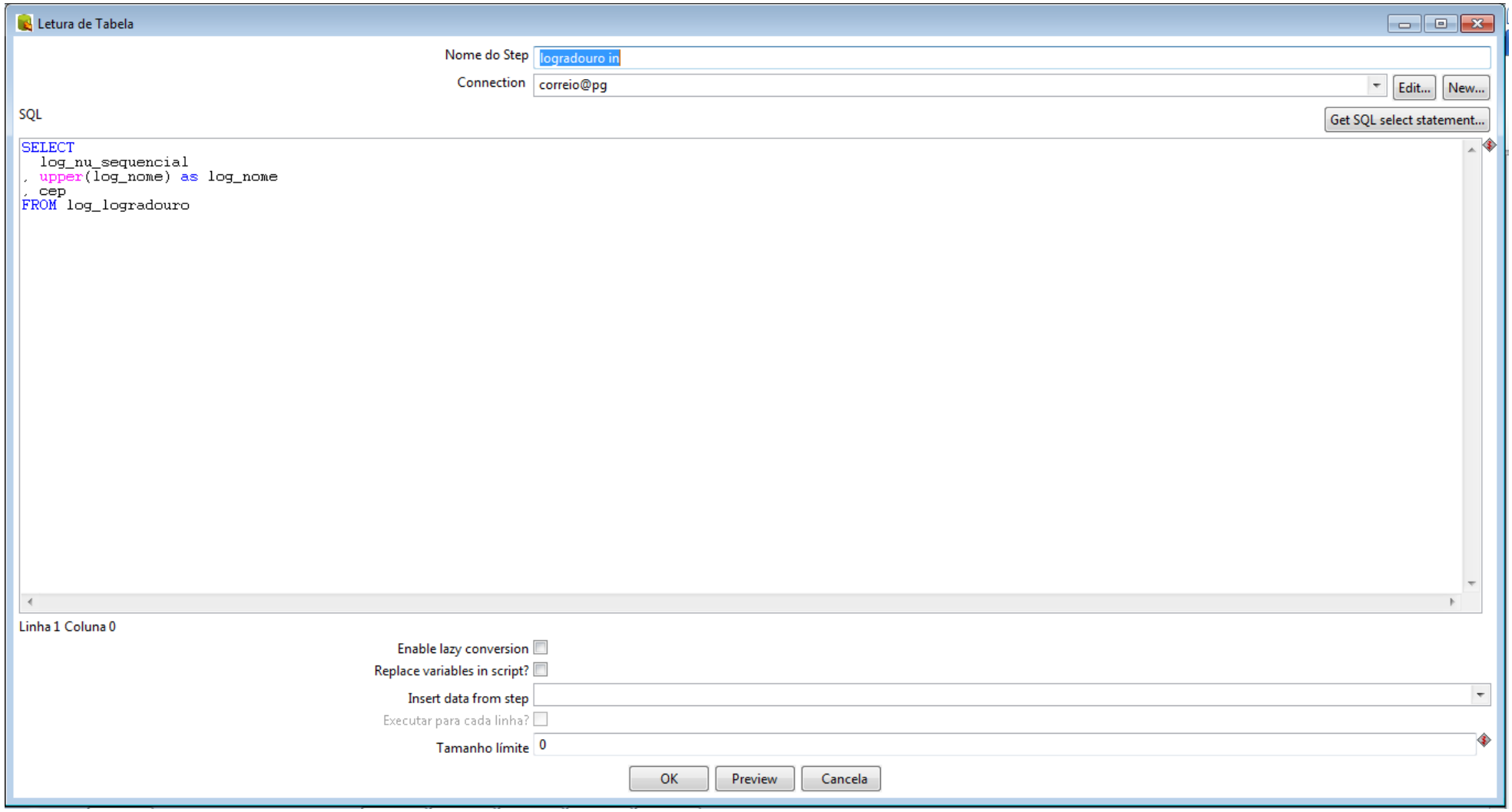


Figura 30 Logradouro in

Como visto na Figura 30, observa-se que para a realização da "Transformação 1 Encontrar CEP por meio de algoritmo de Levenshtein", o select executado na tabela log_logradouro traz os seguintes campos: log_nu_sequencial, log_nome (com comando upper como verificado na Figura 11) e o campo cep.

A transformação continua com o passo do tipo *Fuzzy match,* Figura 31.

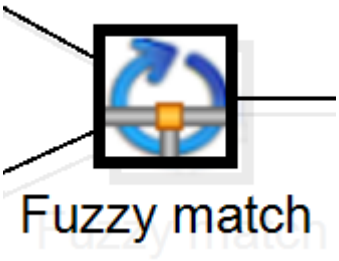


Figura 31 Fuzzy match

Na ferramenta de ETL – PDI o passo do tipo "*Fuzzy match*" é utilizado para encontrar correspondências aproximadas de uma *string* usando algoritmos de similaridade. Lendo um campo de uma fonte principal e encontrando qual valor se aproxima mais da entrada válida.

Ao se abrir o passo "*Fuzzy match*" Figura 32 apresentam-se as opções de "Lookup stream (source)" que trata da fonte de validação onde é inserido o *step* logradouro in e o *field* log_nome que compõem a base com os dados íntegros. Já na opção *Main stream* o parâmetro passado é o *field* que se pretende comparar. Nas opões de configuração seleciona-se o algoritmo a ser utilizado, para esta implementação utiliza-se Levenshtein Distance, além disso, nos campos *minimal* e *maximal value* insere-se a distância máxima e mínima a ser analisada.

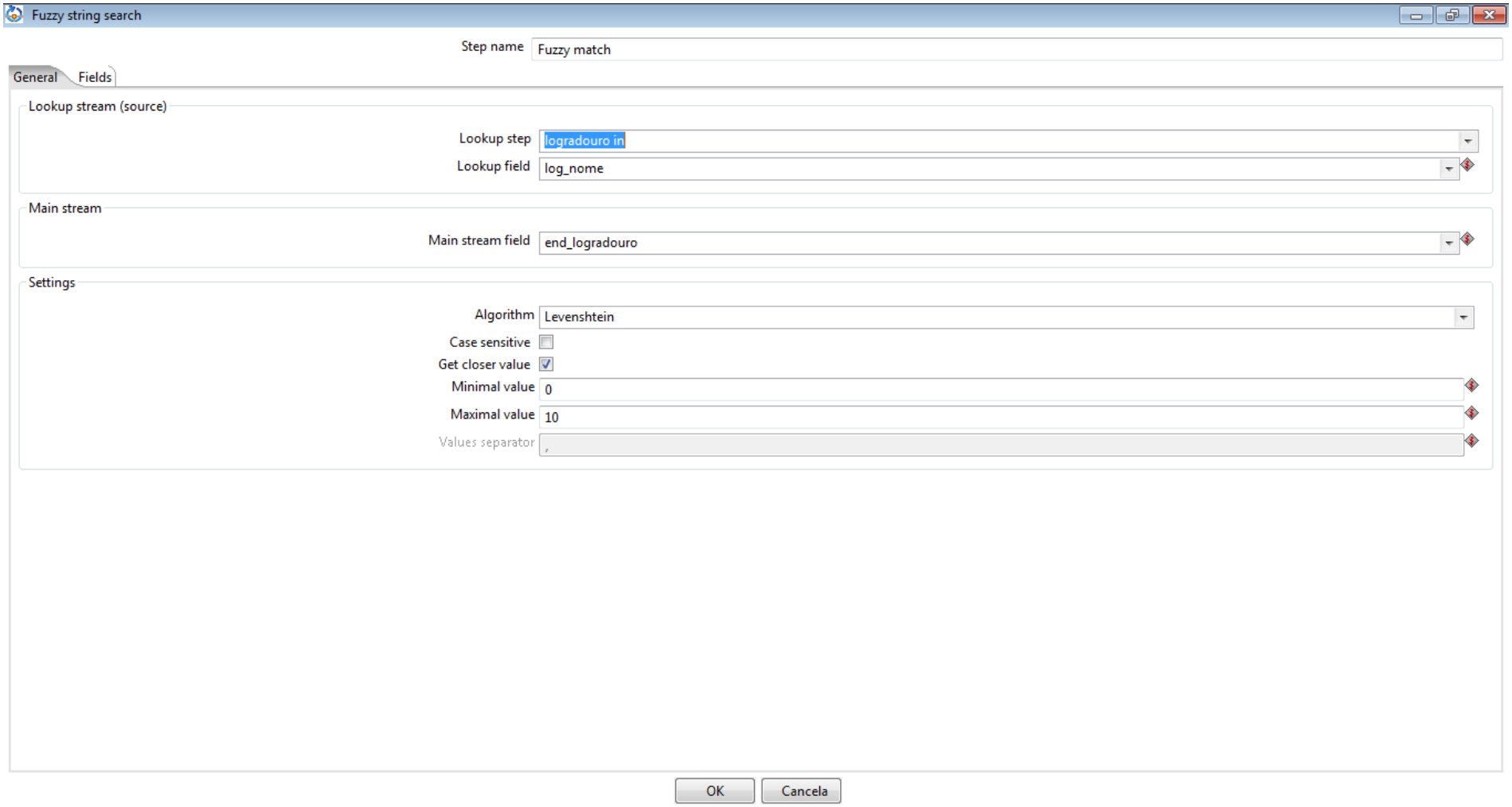


Figura 32 ***Fuzzy match***

A transformação continua com o passo do tipo *Add constants* Figura 33.

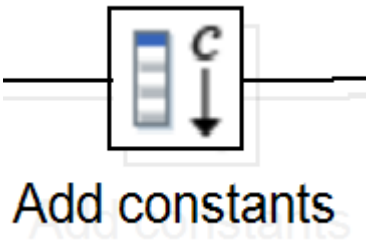


Figura 33 Add constants

Na ferramenta de ETL – PDI o passo do tipo "*Add constants*" é utilizado para adicionar um ou mais campos a tabela transformada.

Ao se abrir o passo "*Add constants*" Figura 34 insere-se um campo de nome passo com valor igual a 1 para identificar dentro da tabela de saída as alterações realizadas pela Transformação 1 - Encontrar CEP por meio de algoritmo de Levenshtein.

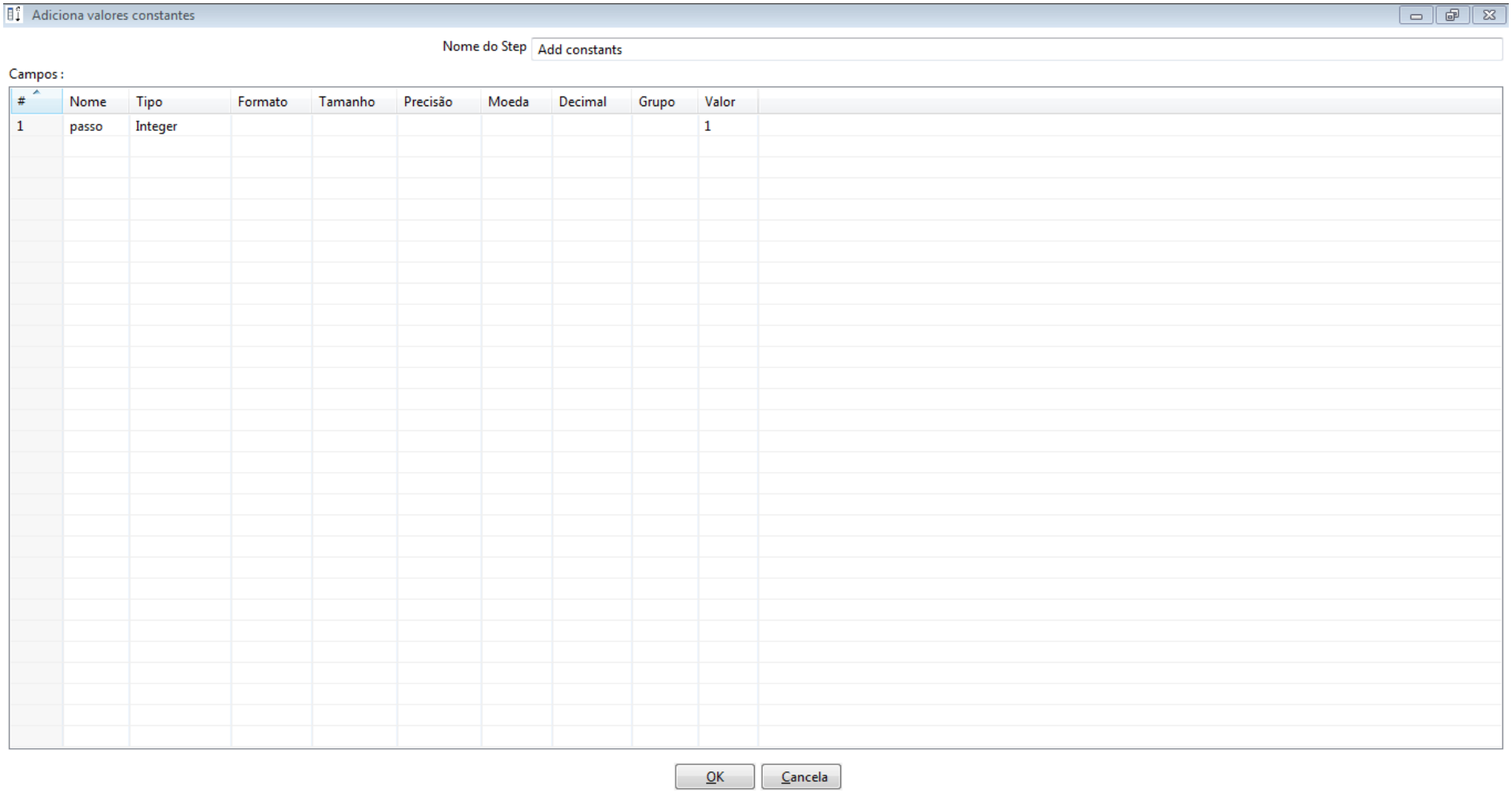


Figura 34 Add constants

A transformação continua com o passo do tipo *Table output* – out resultado Figura 35.

Figura 35 Table output - out resultado

Na ferramenta de ETL – PDI o passo do tipo "*Table output*" é utilizado para criação ou alteração de tabela de saída aonde serão gravados os dados da transformação

Ao se abrir o passo "*Table output*" Figura 36 e Figura 37 insere-se ou altera-se o nome do passo, o nome da conexão, o nome do esquema e o nome da tabela. Além disso, verifica-se o tamanho do commit e se a tabela deve ser truncada antes da inicialização, na segunda aba do table output, a aba *Database fields*, Figura 37, verifica-se a estrutura da tabela que será gerada, bem como, todos os seus campos correspondentes.

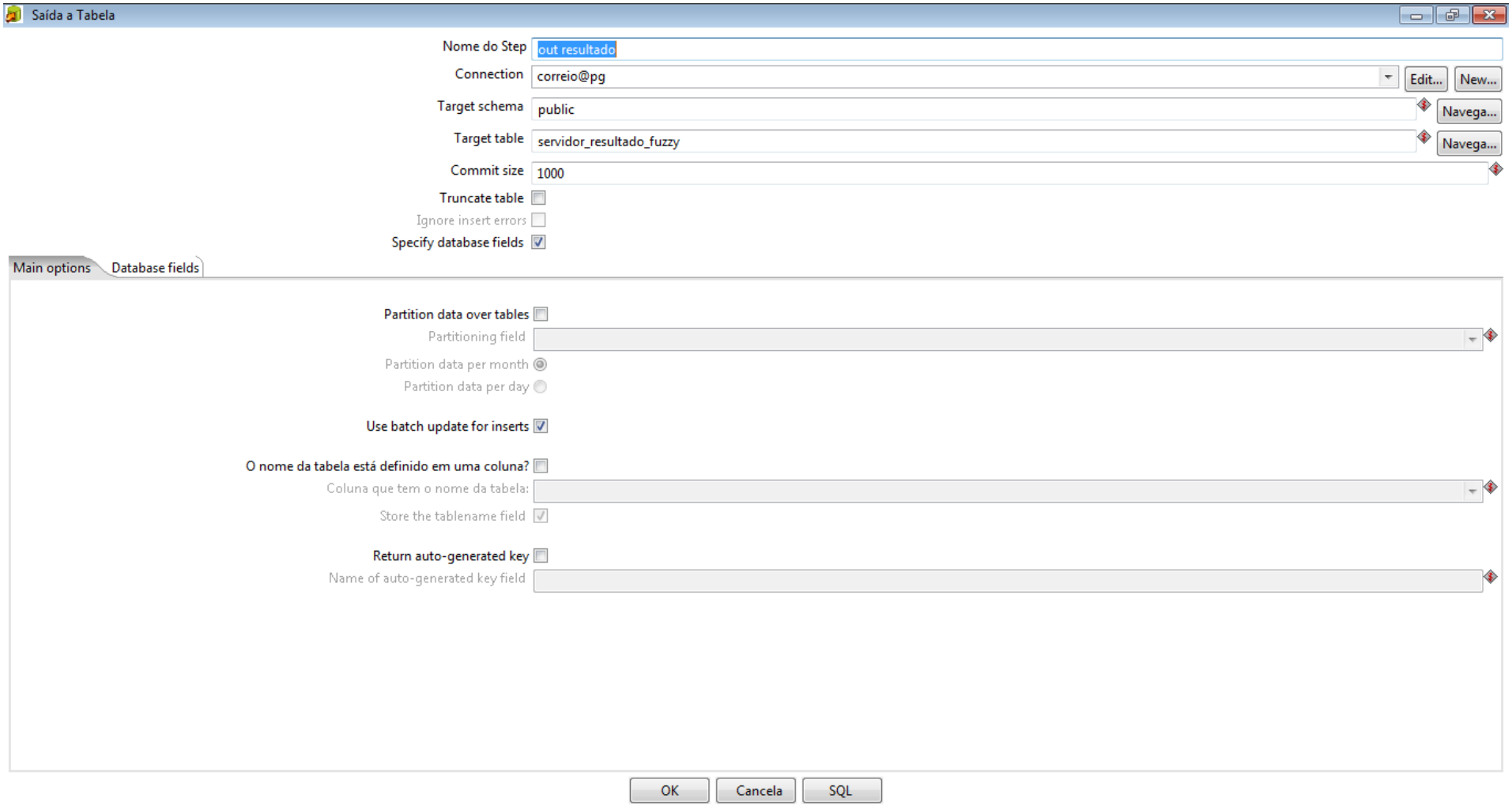


Figura 36 out resultado - aba 1

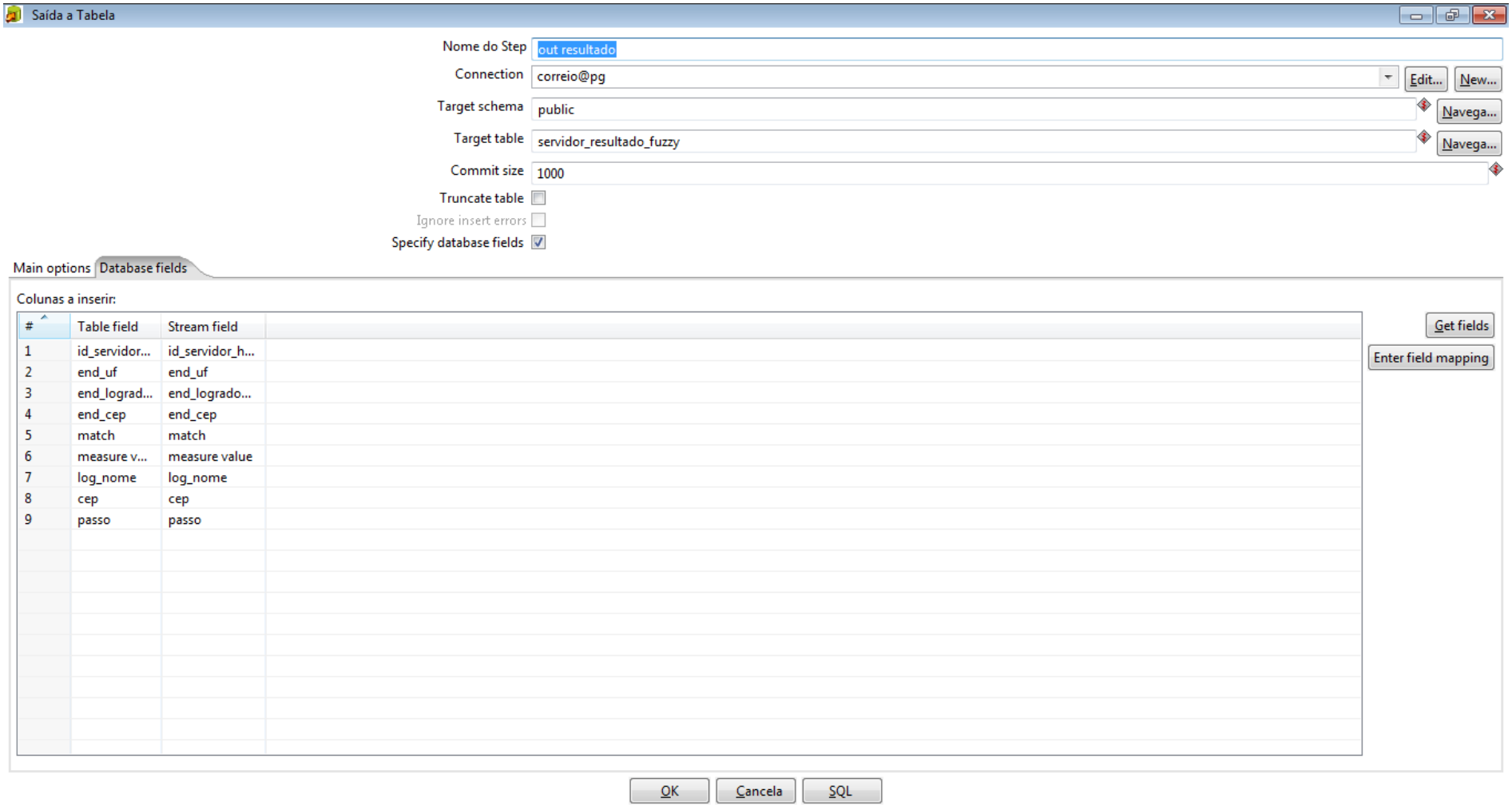


Figura 37 Out resultado - aba 2

A Figura 38 traz toda “Transformação 2 - Encontrar Logradouro por meio de algoritmo de Levenshtein” realizada por meio da ferramenta de ETL – PDI.

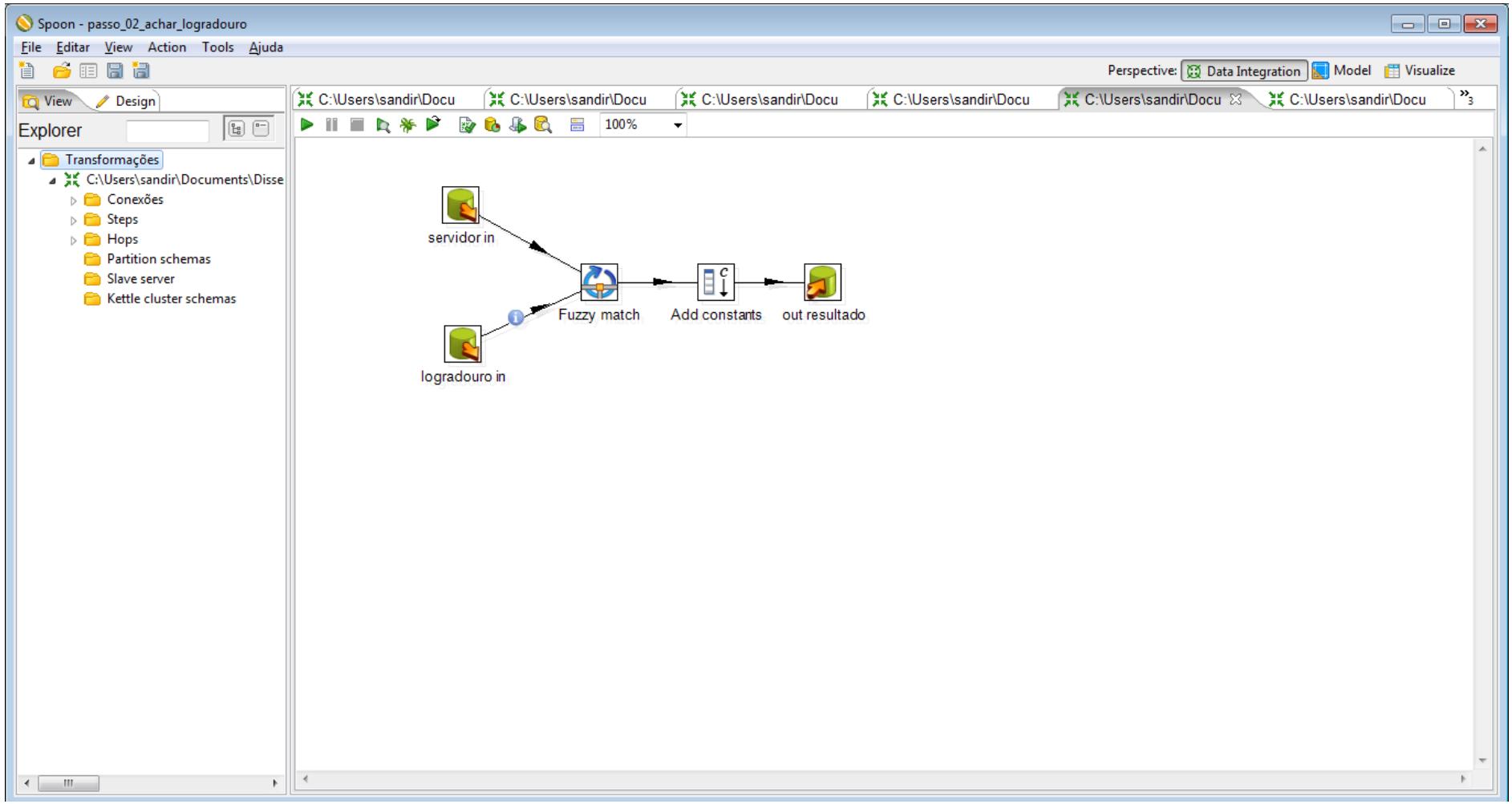


Figura 38 – Transformação 2 - Encontrar Logradouro por meio de algoritmo de Levenshtein

A transformação inicia-se com o passo do tipo Table Input servidor in como mostrado na Figura 27.

Ao se abrir o passo servidor in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados gerado pela extração do arquivo espelho do SIAPE na tabela siape_servidor como verificado na Figura 39.

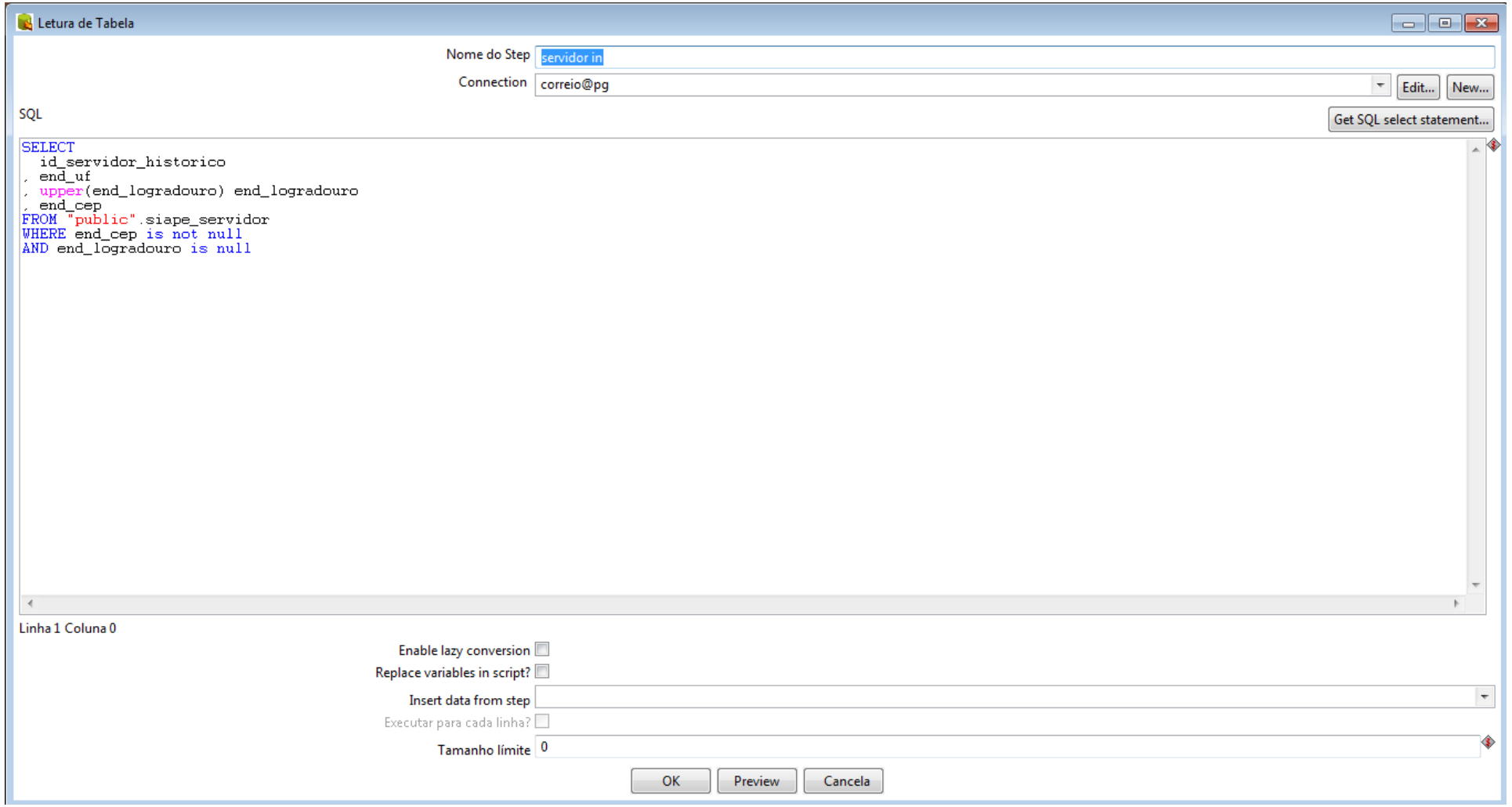


Figura 39 Servidor in passo 2

Como visto na Figura 39, utiliza-se a entrada de dados para a realização da “Transformação 2 - Encontrar Logradouro por meio de algoritmo de Levenshtein”, o comando SQL executado na tabela siape_servidor traz os seguintes campos: id_servidor_historico (identificador único do servidor), end_uf (UF do endereço do servidor), end_logradouro em caixa alta para que possa ser comparada com a base do DNE e por fim o campo end_cep (traz o CEP cadastrado para o servidor).

O comando SQL de seleção realizado, também condiciona aos registros que possuírem o campo end_cep com algum valor preenchido e o campo end_logradouro sem nenhum valor inserido.

A transformação continua com o passo do tipo *Table Input* logradouro in como mostrado na Figura 29.

Ao se abrir o passo logradouro in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados do DNE dentro da tabela log_logradouro, como visualizado na Figura 30.

Como visto na Figura 30, reutiliza-se a entrada de dados para a realização da “Transformação 2 - Encontrar Logradouro por meio de algoritmo de Levenshtein”, o comando SQL de seleção executado na tabela log_logradouro traz os seguintes campos: log_nu_sequencial, log_nome (com comando upper como verificado na Figura 28) e o campo cep.

A transformação continua com o passo do tipo *Fuzzy match*, ver Figura 31.

Ao se abrir o passo "*Fuzzy match*" da "Transformação 2 - Encontrar Logradouro por meio de algoritmo de Levenshtein" Figura 40 apresentam-se as opções de "Lookup stream (source)" que tratam da fonte de validação onde é inserido o *step* logradouro in e o *field* cep que compõem a base com os dados íntegros. Já na opção *Main stream* o parâmetro passado é o *field* que se pretende comparar, neste caso o campo end_cep. Nas opões de configuração seleciona-se o algoritmo a ser utilizado, para esta implementação utiliza-se *Levenshtein Distance*, além disso, nos campos *minimal* e *maximal value* insere-se a distância máxima e mínima a ser analisada.

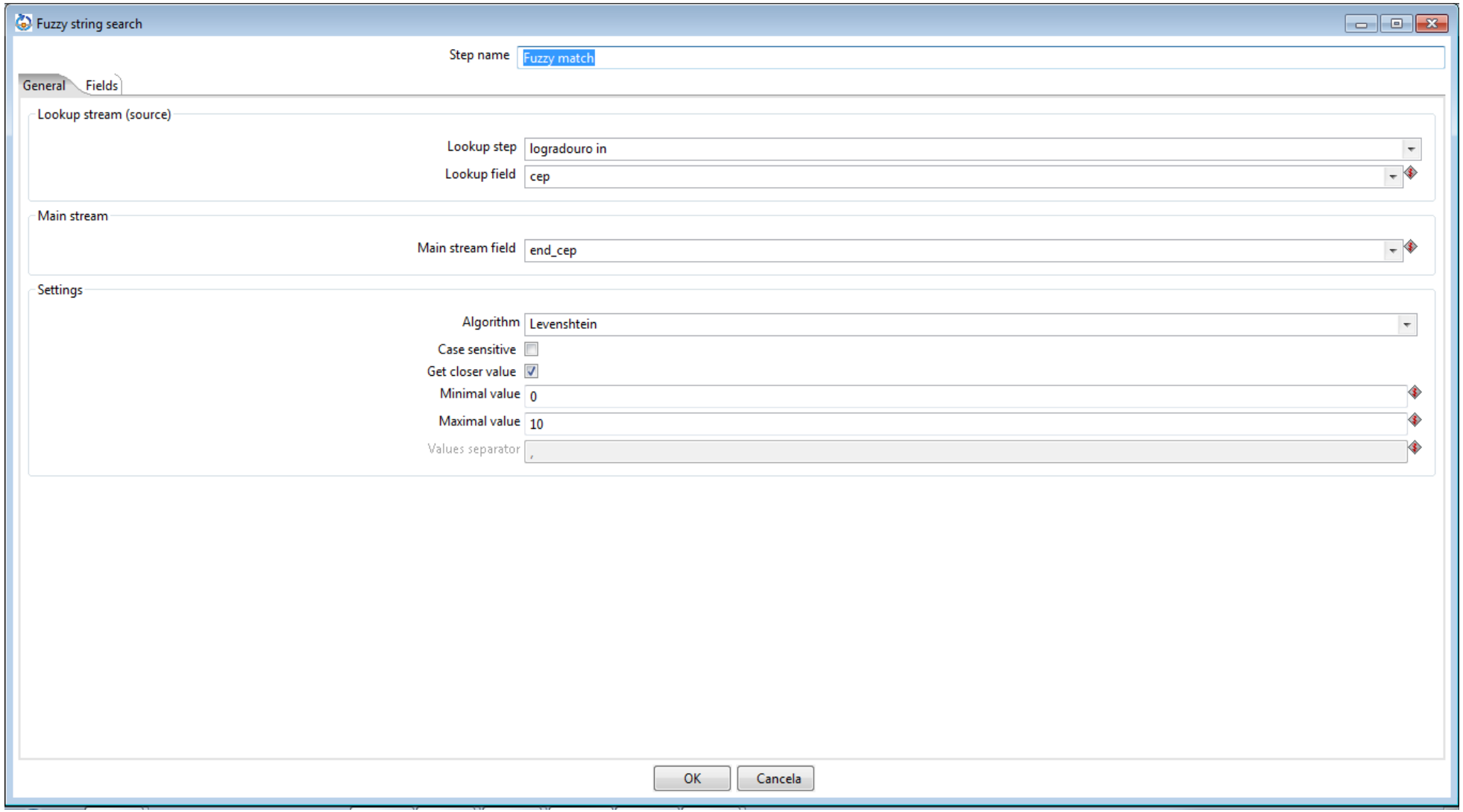


Figura 40 Fuzzy match passo 2

A transformação continua com o passo do tipo *Add constants*, conforme Figura 16.

Ao se abrir o passo "*Add constants*" Figura 41 insere-se um campo de nome passo com valor igual a 2 para identificar dentro da tabela de saída as alterações realizadas pela Transformação 2 - Encontrar Logradouro por meio de algoritmo de Levenshtein.

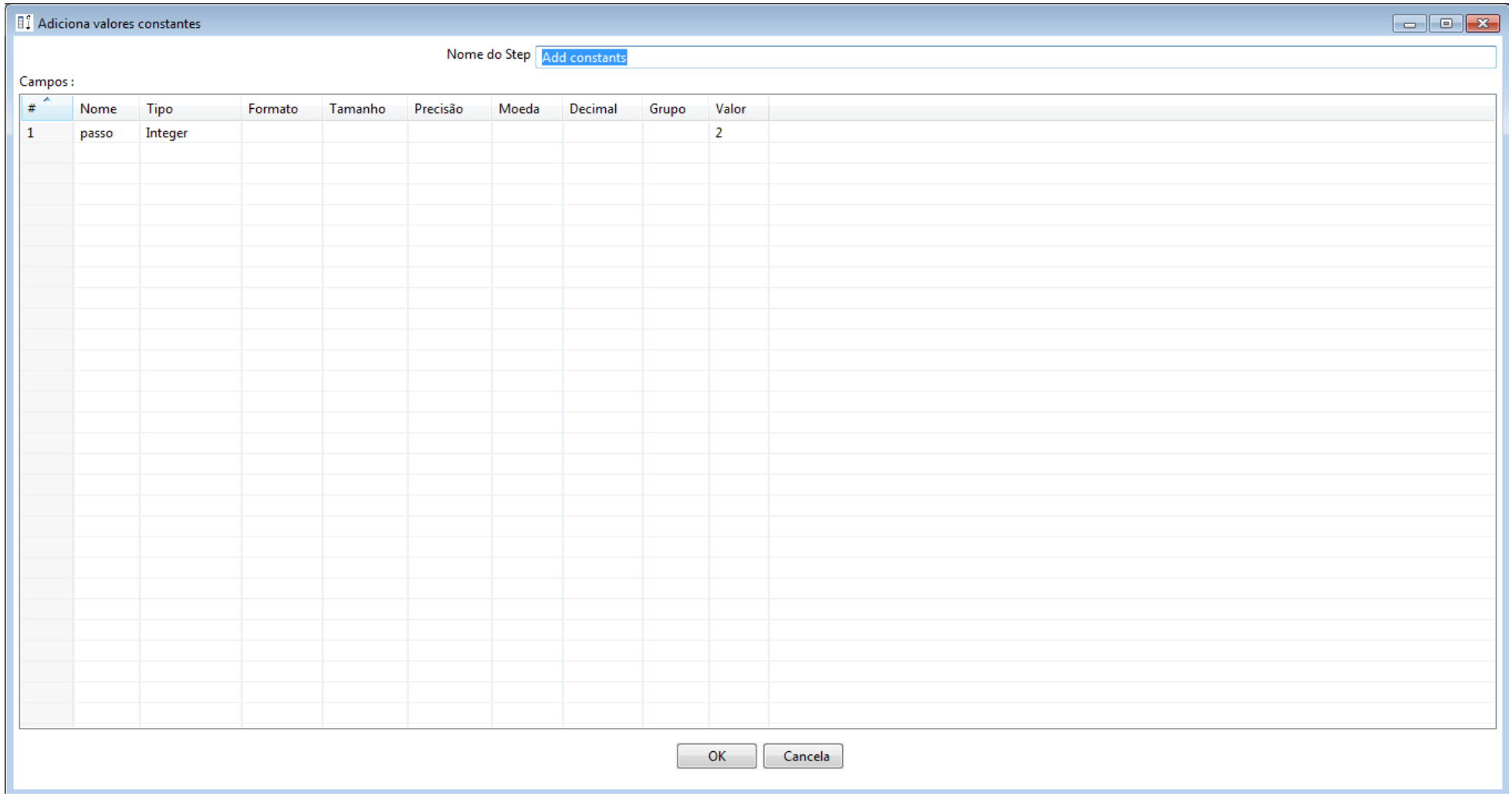


Figura 41 Add constants passo2

A transformação continua com o passo do tipo *Table output* – out resultado, como visto na Figura 35.

Ao se abrir o passo "*Table output*", ver Figura 36 e Figura 37, insere-se ou altera-se o nome do passo, o nome da conexão, o nome do esquema e o nome da tabela. Além disso, verifica-se o tamanho do commit e se a tabela deve ser truncada antes da inicialização, na segunda aba do table output, a aba *Database fields*, ver Figura 37, verifica-se a estrutura da tabela que será gerada, bem como, todos os seus campos correspondentes.

A Figura 42 traz toda "Transformação 3 - Completar CEP genérico por meio de algoritmo de Levenshtein" realizada por meio da ferramenta de ETL – PDI.

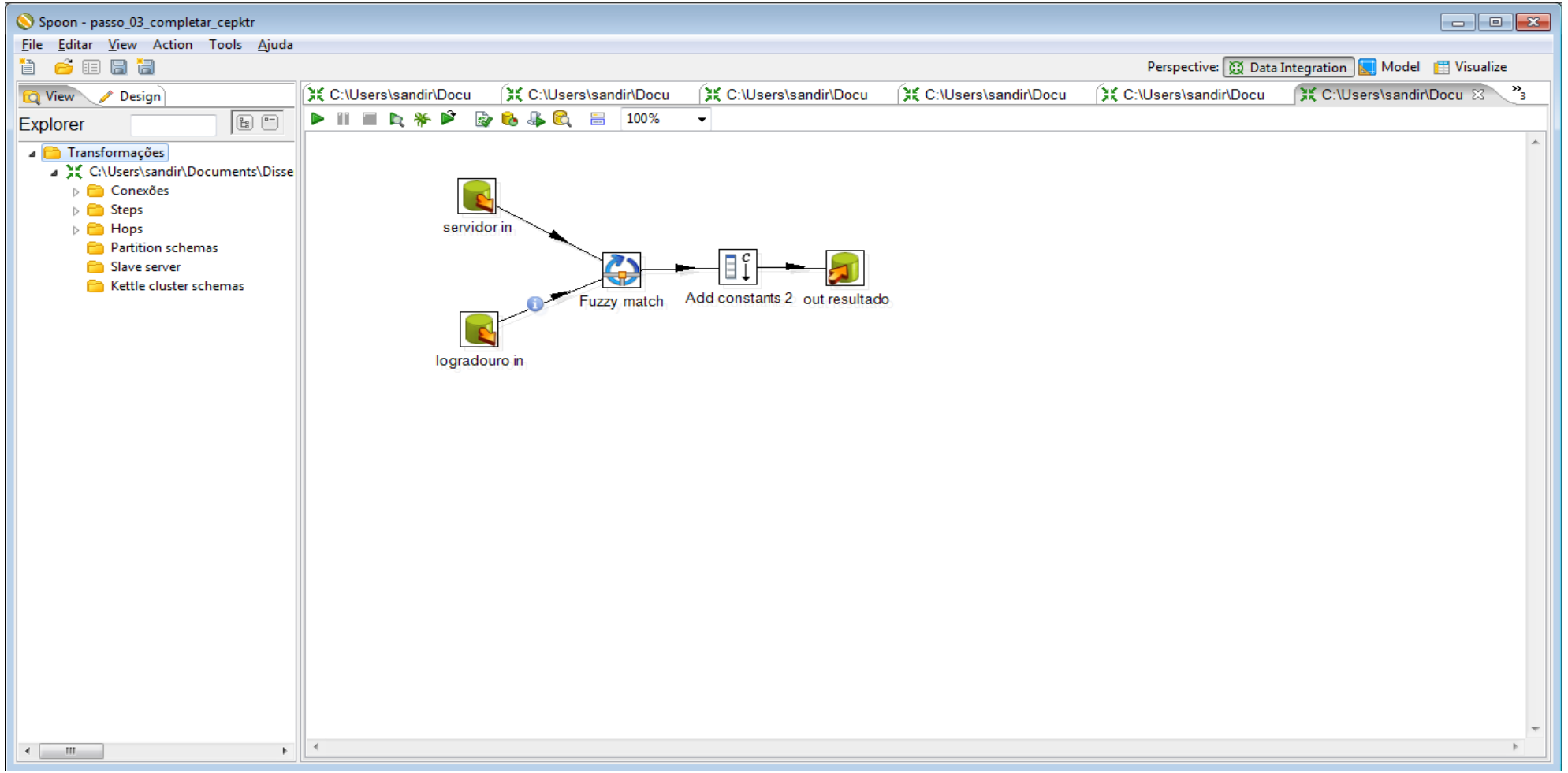


Figura 42 – Transformação 3 - Completar CEP genérico por meio de algoritmo de Levenshtein

A transformação inicia-se com o passo do tipo Table Input servidor in como mostrado na Figura 26.

Ao se abrir o passo servidor in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados gerado pela extração do arquivo espelho do SIAPE na tabela siape_servidor Figura 43.

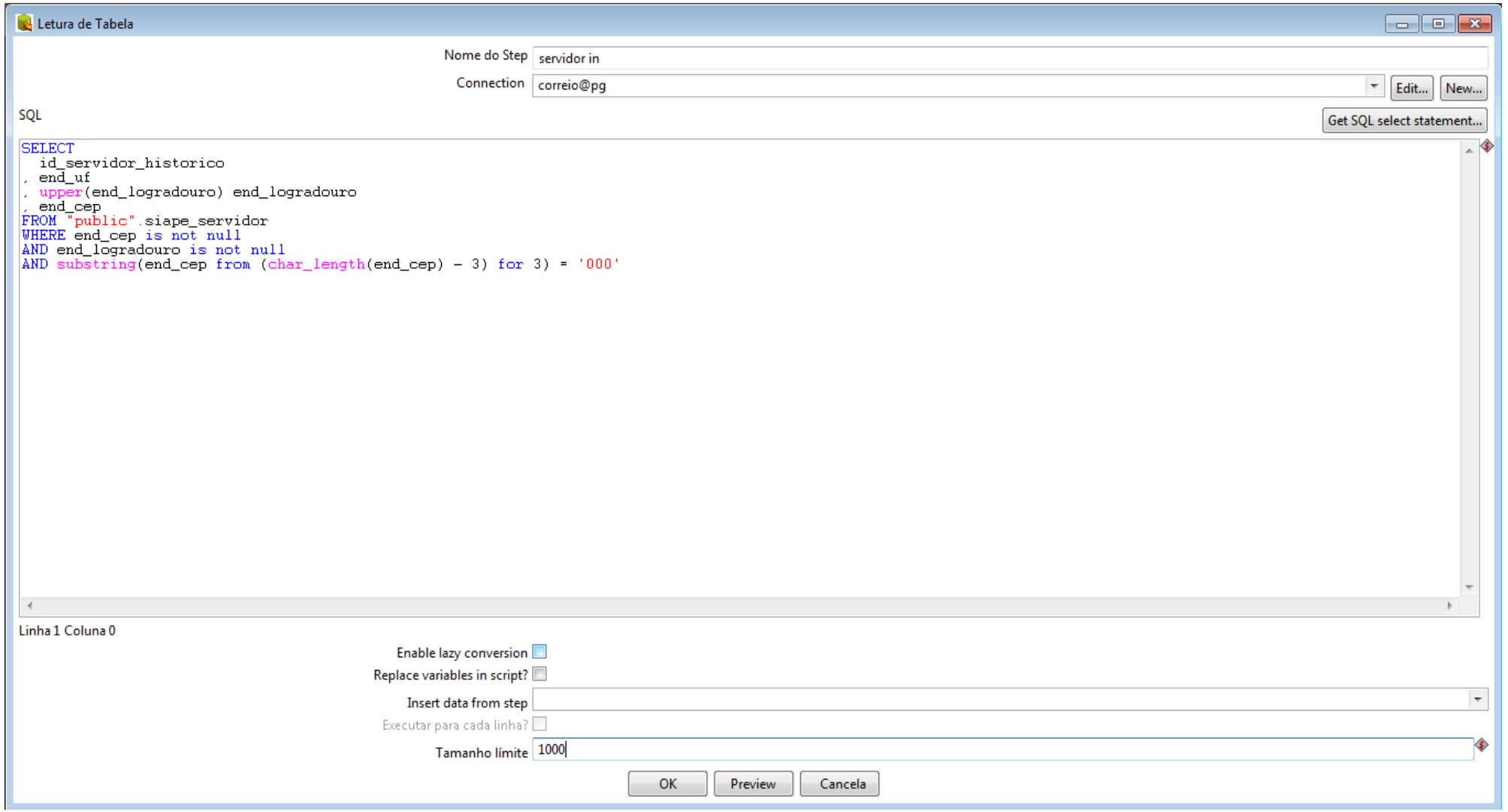


Figura 43 Servidor in passo 3

Como visto na Figura 43, utiliza-se a entrada de dados para a realização da "Transformação 3 - Completar CEP genérico por meio de algoritmo de Levenshtein", o comando SQL executado na tabela siape_servidor traz os seguintes campos: id_servidor_historico(identificador único do servidor), end_uf (UF do endereço do servidor), end_logradouro em caixa alta para que possa ser comparada com a base do DNE e por fim o campo end_cep (traz o CEP cadastrado para o servidor). Para execução deste passo, devido a um excessivo tempo de processamento, foi necessário aplicar uma limitação de 1000 registros, preenchendo o campo tamanho limite.

O comando SQL de seleção realizado, também condiciona aos registros que possuírem o campo end_cep e o campo end_logradouro com algum valor inserido além de só trazer os valores do campo end_cep que possuírem os três últimos dígitos com valor igual a zero.

A transformação continua com o passo do tipo *Table Input* logradouro in como mostrado na Figura 29.

Ao se abrir o passo logradouro in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados do DNE dentro da tabela log_logradouro, como visualizado na Figura 30.

Como visto na Figura 30, reutiliza-se a entrada de dados para a realização da "Transformação 3 - Completar CEP genérico por meio de algoritmo de Levenshtein", o comando SQL de seleção executado na tabela log_logradouro traz os seguintes campos: log_nu_sequencial, log_nome (com comando upper como verificado na Figura 28) e o campo cep.

A transformação continua com o passo do tipo *Fuzzy match*, ver Figura 31.

Ao se abrir o passo "*Fuzzy match*" da "Transformação 3 - Completar CEP genérico por meio de algoritmo de Levenshtein" Figura 44 apresentam-se as opções de "Lookup stream (source)" que tratam da fonte de validação onde é inserido o *step* logradouro in e o *field* log_nome que compõem a base com os dados íntegros. Já na opção *Main stream* o parâmetro passado é o *field* que pretende-se comparar, neste caso o campo end_logradouro. Nas opões de configuração seleciona-se o algoritmo a ser utilizado, para esta implementação utiliza-se *Levenshtein Distance*, além disso, nos campos *minimal* e *maximal value* insere-se a distância máxima e mínima a ser analisada.

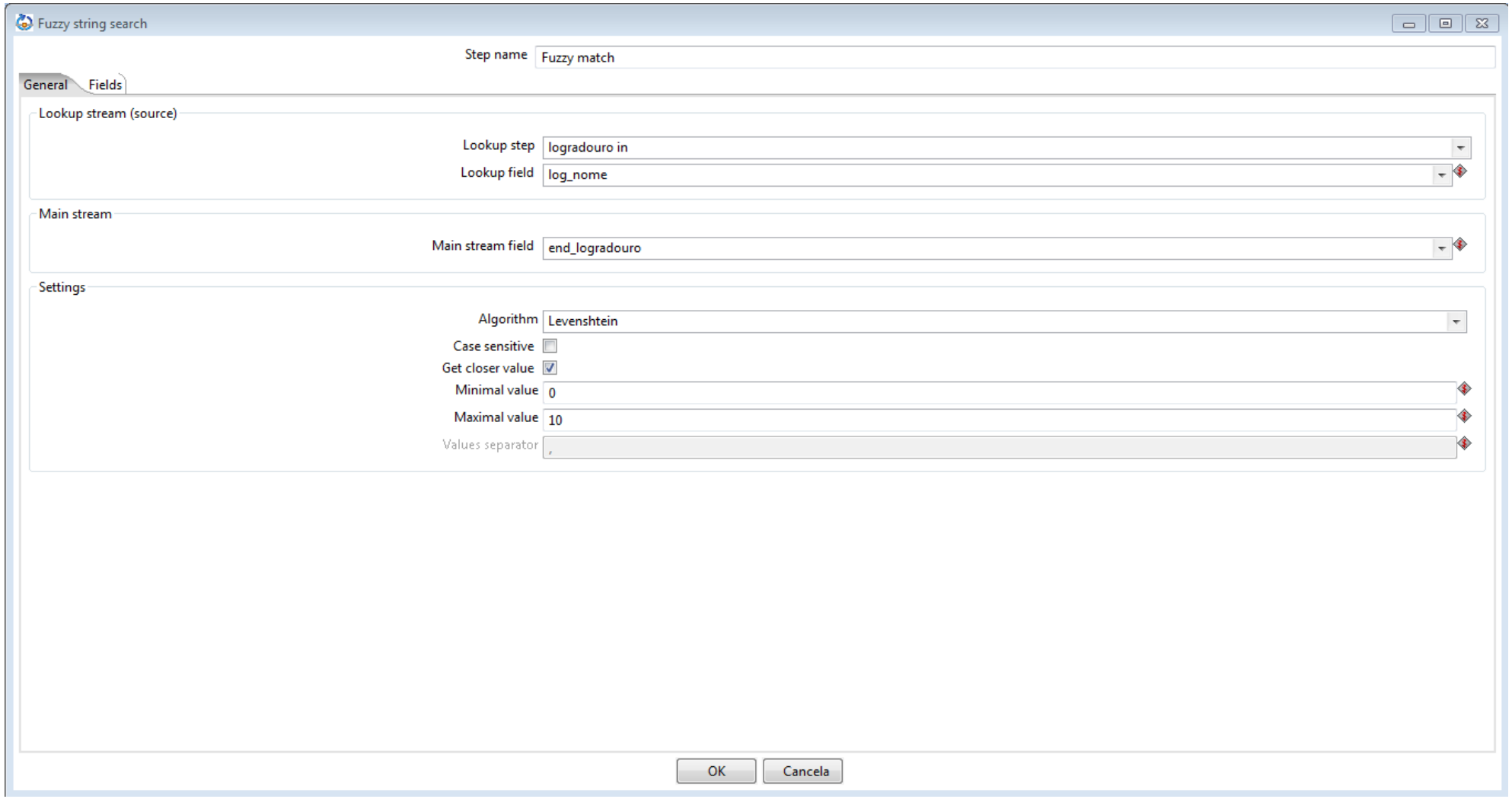


Figura 44 Fuzzy match passo 3

A transformação continua com o passo do tipo *Add constants*, conforme Figura 33.

Ao se abrir o passo "*Add constants*" Figura 45 insere-se um campo de nome passo com valor igual a 3 para identificar dentro da tabela de saída as alterações realizadas pela Transformação 3 - Completar CEP genérico por meio de algoritmo de Levenshtein.

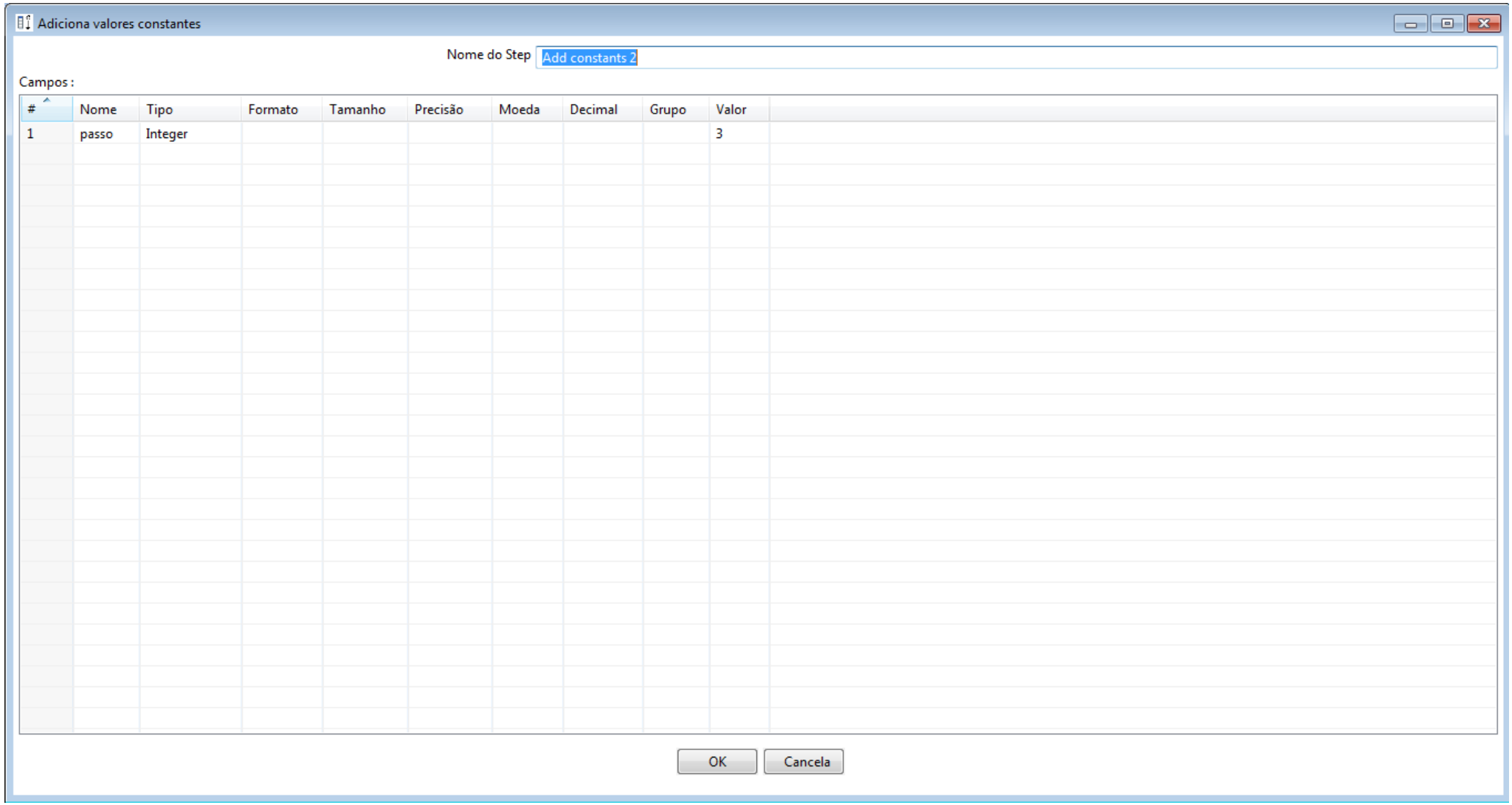


Figura 45 Add constants passo 3

A transformação continua com o passo do tipo *Table output* – out resultado, como visto na Figura 35.

Ao se abrir o passo "*Table output*", ver Figura 36 e Figura 37, inserem-se ou altera-se o nome do passo, o nome da conexão, o nome do esquema e o nome da tabela. Além disso, verifica-se o tamanho do commit e se a tabela deve ser truncada antes da inicialização, na segunda aba do table output, a aba *Database fields*, ver Figura 37, verifica-se a estrutura da tabela que será gerada, bem como, todos os seus campos correspondentes.

ANEXO B – APLICAÇÃO DA ADAPTAÇÃO DO COEFICIENTE DE DICE

A adaptação do coeficiente de *Dice* passa por uma nova maneira de tratar frases inteiras dos logradouros que se deseja comparar, sob esta nova perspectiva a frase do logradouro é separada em palavras, ou seja cada palavra separada por espaço é considerada de forma individual, posteriormente é calculada o coeficiente de similaridade de cada string, depois somam-se os coeficientes e divide-se tudo pelo número de palavras do conjunto.

A aplicação do algoritmo de coeficiente de *Dice* na ferramenta de ETL – PDI necessita de inserção de componentes externos como visto na Figura 46.

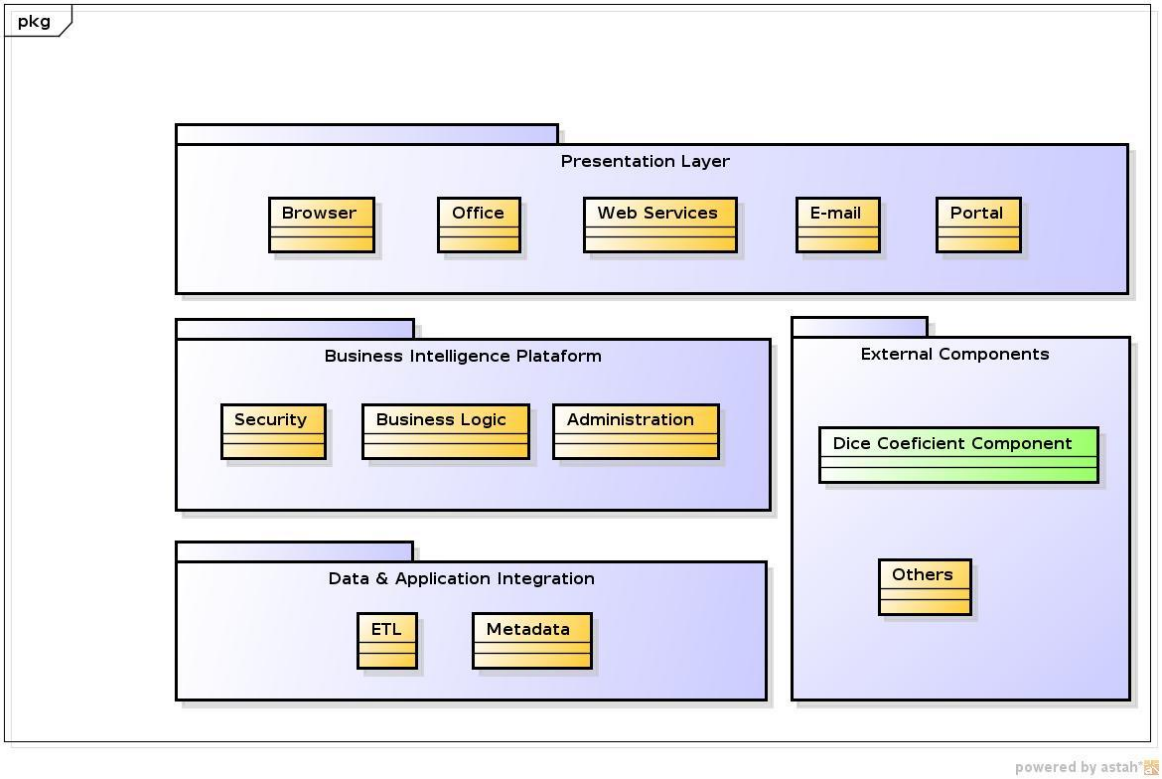


Figura 46 Encapsulamento das aplicações da plataforma Pentaho

Abaixo se apresentam as imagens dos processos de ETL implementados para realização do estude de caso, utilizando-se da adaptação do coeficiente de Dice configurado para

encontrar e tratar dois casos de coeficiencia de similaridade um com no mínimo 85% de semelhança e outro com um mínimo de 50% de similaridade:

A Figura 47traz toda “Transformação 1 - Encontrar CEP por meio do coeficiente de Dice” realizada por meio da ferramenta de ETL – PDI.

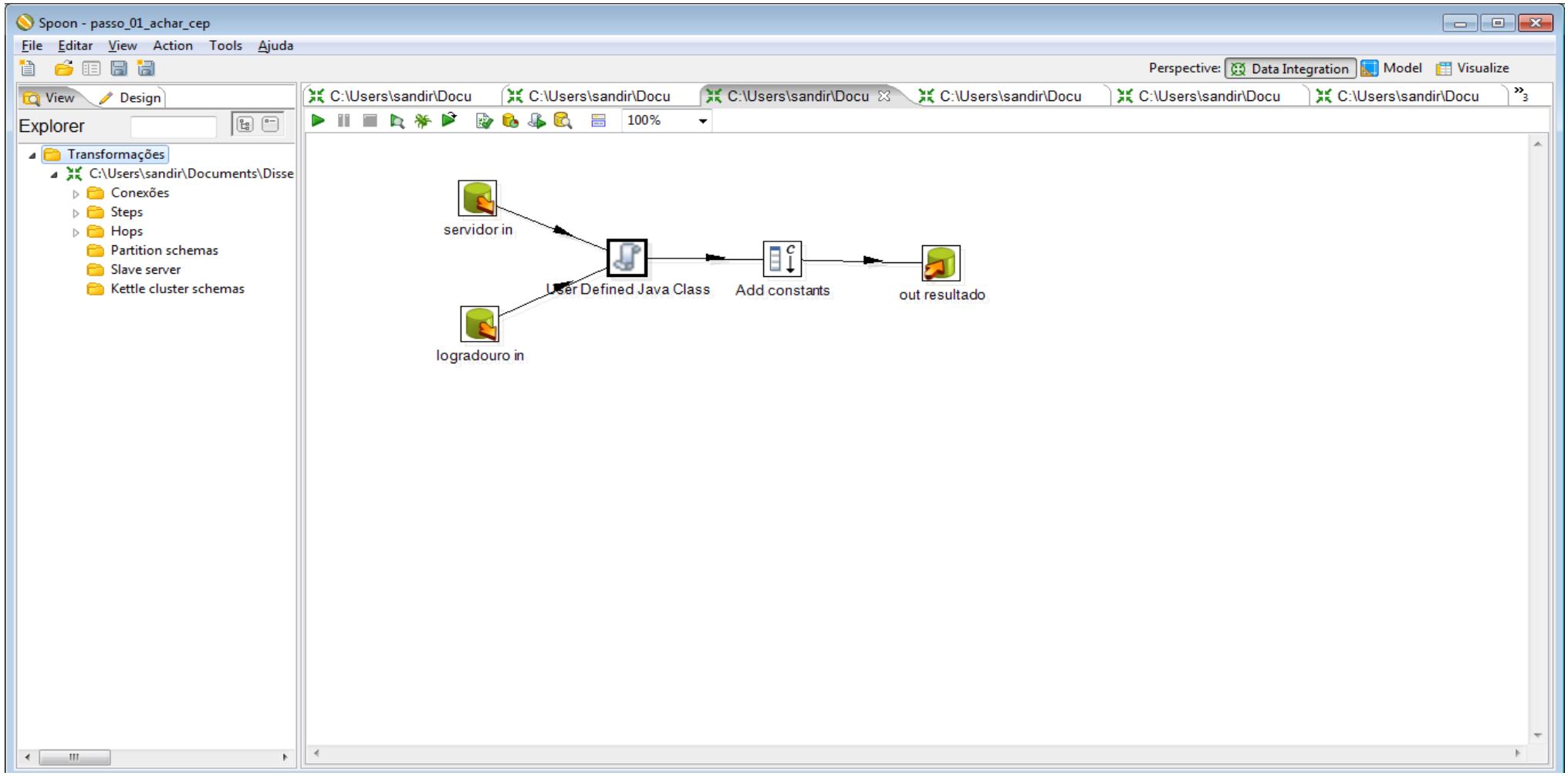


Figura 47 – Transformação 1 - Encontrar CEP por meio do coeficiente de Dice

A transformação inicia-se com o passo do tipo *Table Input* servidor in, como visto na Figura 26.

Ao se abrir o passo servidor in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados gerado pela extração do arquivo espelho do SIAPE na tabela siape_servidor, ver Figura 27.

Como visto na Figura 27, observa-se que para a realização da "Transformação 1 - Encontrar CEP por meio do coeficiente de Dice", o comando SQL executado na tabela siape_servidor traz os seguintes campos: id_servidor_historico (identificador único do servidor), end_uf (UF do endereço do servidor), end_logradouro em caixa alta para que possa ser comparada com a base do DNE e por fim o campo end_cep (traz o CEP cadastrado para o servidor).

Verificou-se a necessidade de utilização do comando "*upper*" no intuito de transformar todos os caracteres de um campo do tipo "*string*" em caracteres de caixa alta, como visto na Figura 28 exibe-se uma consulta realizada sem o comando "*upper*" onde pode ser verificado em destaque um endereço de logradouro em caixa baixa.

O comando SQL de seleção realizado, também condiciona aos registros que não possuírem o campo end_cep sem nenhum valor inserido e o campo end_logradouro com algum valor inserido valor inserido.

A transformação continua com o passo do tipo *Table Input* logradouro in, ver Figura 29.

Ao se abrir o passo logradouro in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados do DNE dentro da tabela log_logradouro, ver Figura 30.

Como visto na Figura 30, observa-se que para a realização da "Transformação 1 - Encontrar CEP por meio do coeficiente de Dice", o comando SQL de seleção executado na tabela log_logradouro traz os seguintes campos: log_nu_sequencial, log_nome (com comando *upper* como verificado na Figura 11) e o campo cep.

A transformação continua com o passo do tipo *User Defined Java Class*, Figura 48.

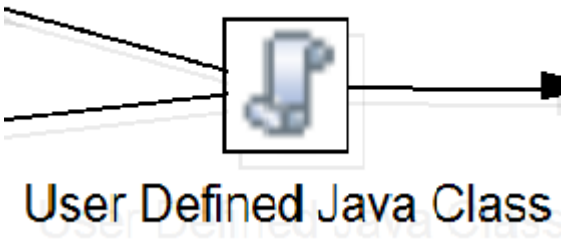


Figura 48 Classe Java passo 1

Na ferramenta de ETL – PDI o passo do tipo "*User Defined Java Class*" é utilizado para utilizar classes de código na linguagem de programação Java.

Ao se abrir o passo "*User Defined Java Class*" Figura 49 (verificar a que figura se refere) a função apresenta o processo que foi utilizado para realizar uma chamada de importação da função da similaridade Dice com os parâmetros de entrada.

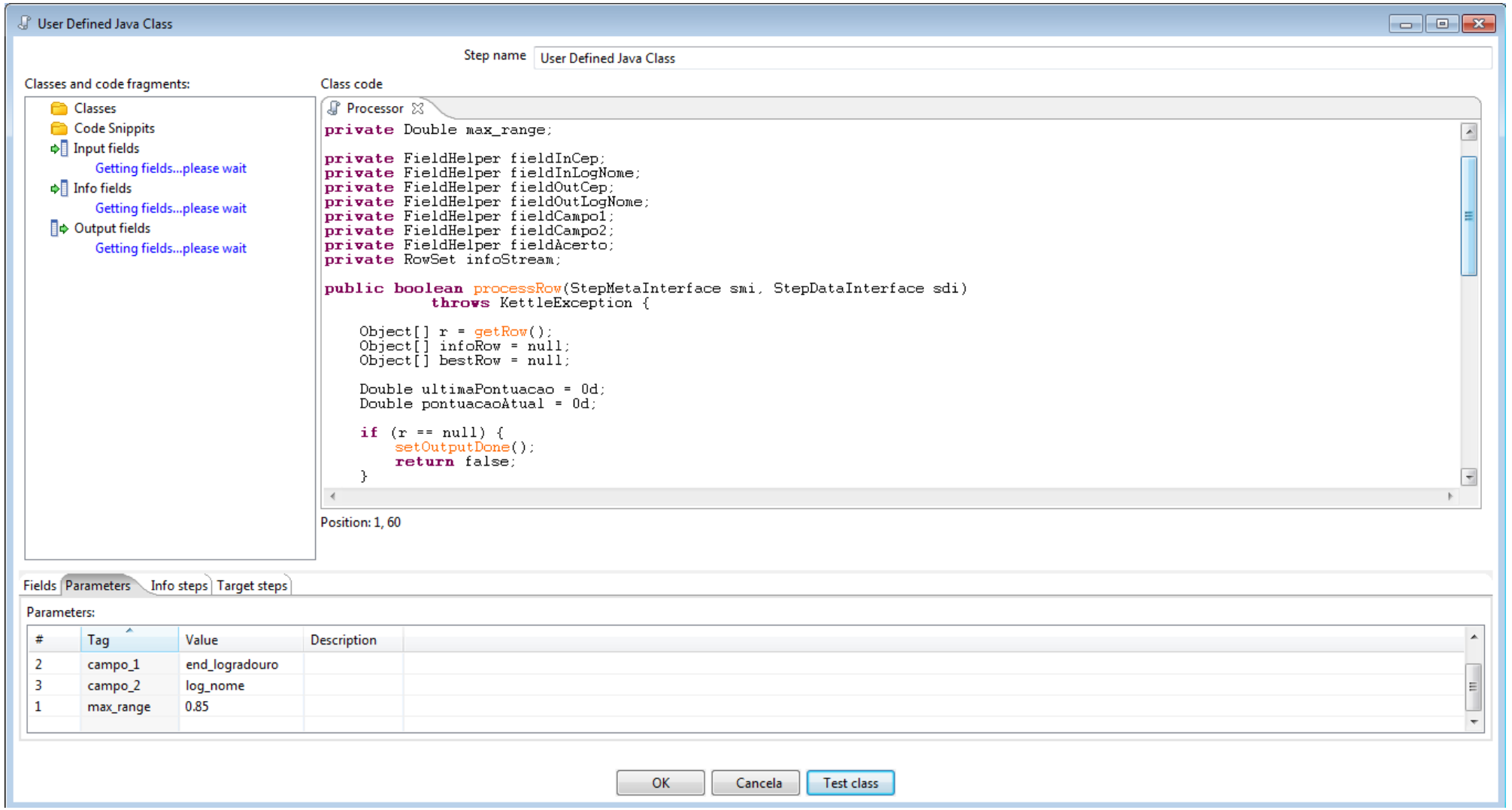


Figura 49 Java Class passo 1

A transformação continua com o passo do tipo *Add constants*, ver Figura 33.

Ao se abrir o passo "*Add constants*", ver Figura 50, insere-se um campo de nome passo com valor igual a 1 para identificar dentro da tabela de saída as alterações realizadas pela Transformação 1 - Encontrar CEP por meio do coeficiente de Dice.

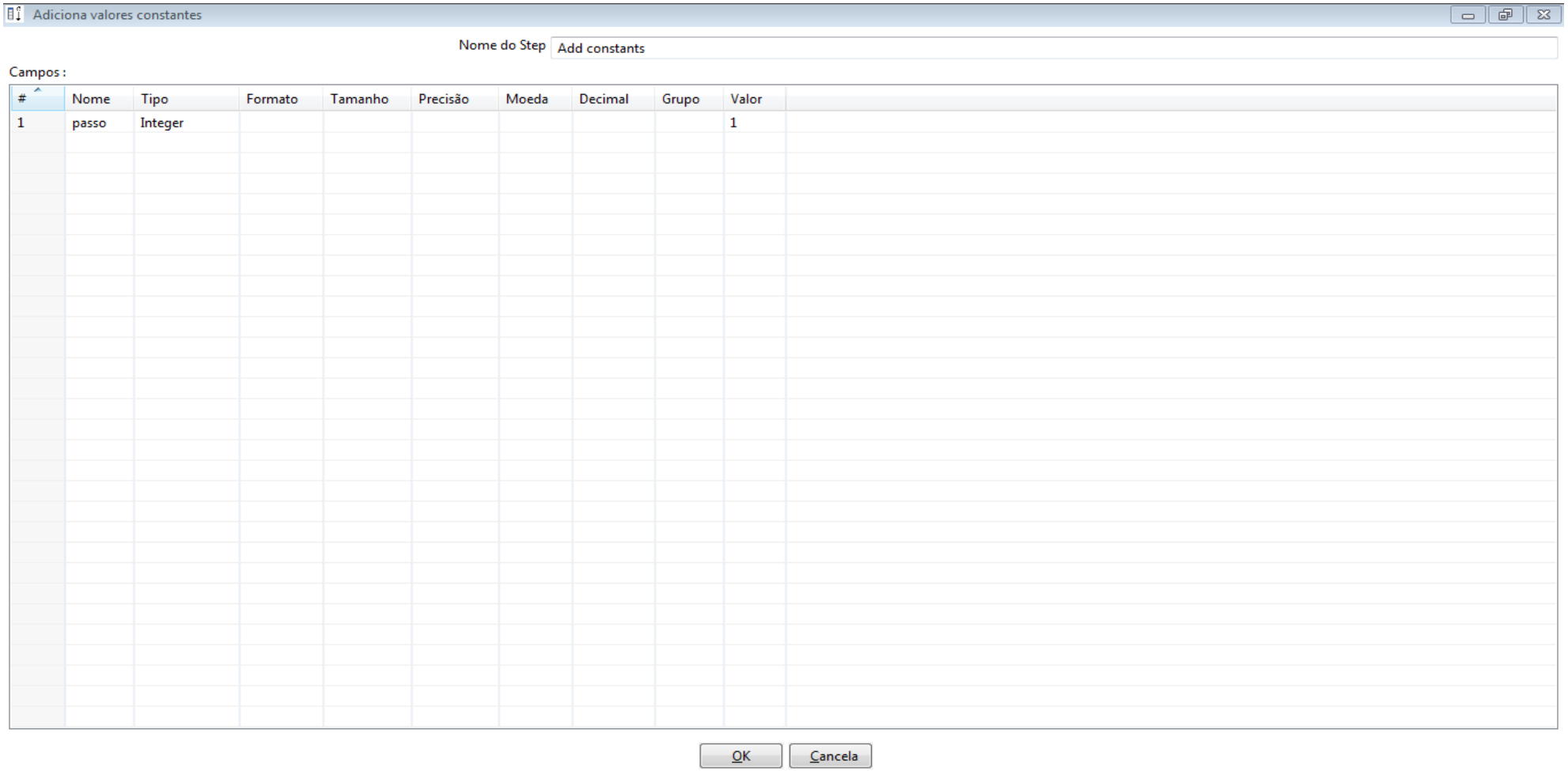


Figura 50 Add constants

A transformação continua com o passo do tipo *Table output* – out resultado, ver Figura 35.

Ao se abrir o passo "*Table output*" Figura 51 e Figura 52 inserem-se ou altera-se o nome do passo, o nome da conexão, o nome do esquema e o nome da tabela. Além disso, verifica-se o tamanho do commit e se a tabela deve ser truncada antes da inicialização, na segunda aba do table output, a aba *Database fields*, Figura 52, verifica-se a estrutura da tabela que será gerada, bem como, todos os seus campos correspondentes.

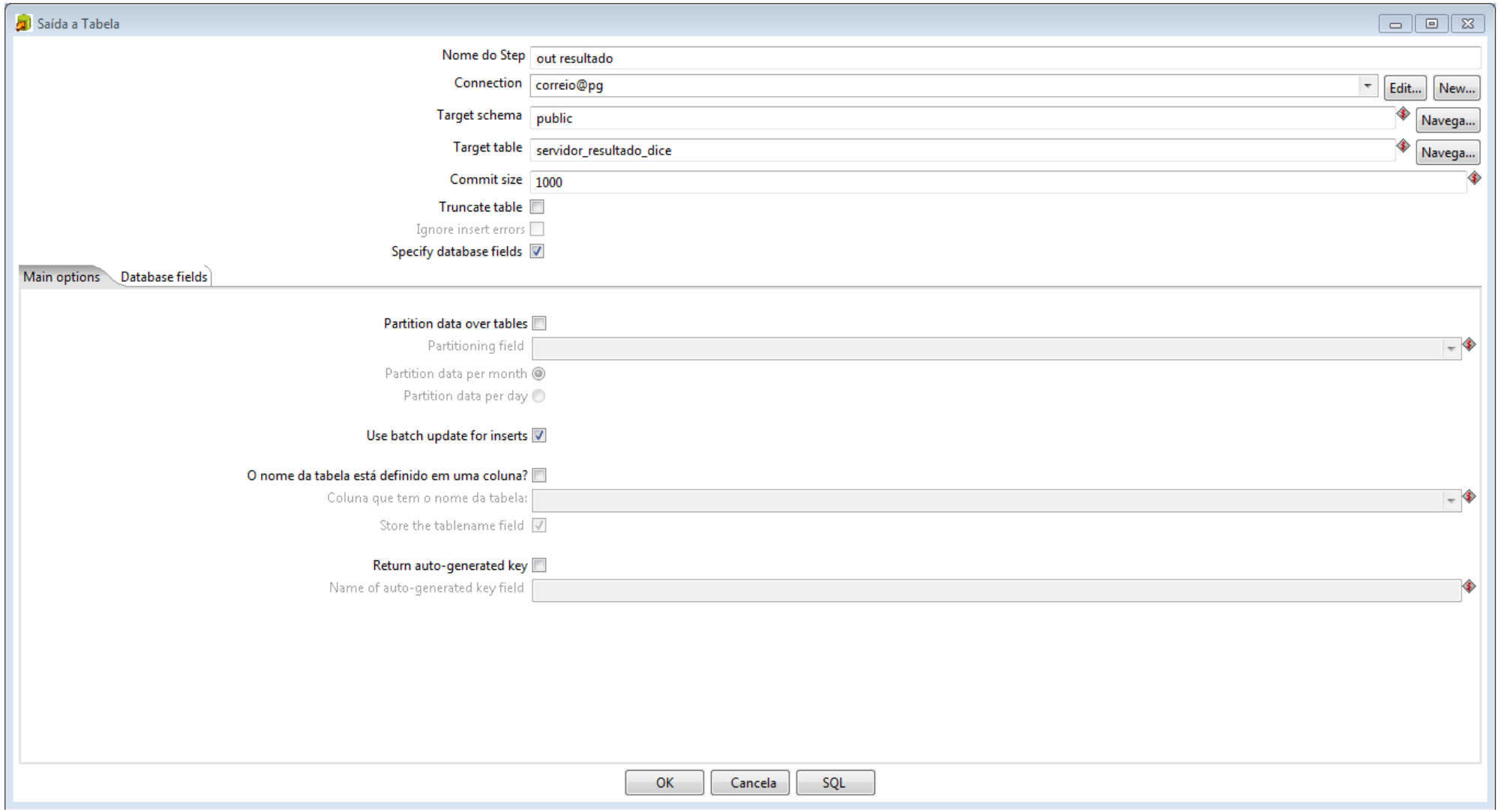


Figura 51 Out resultado - aba 1

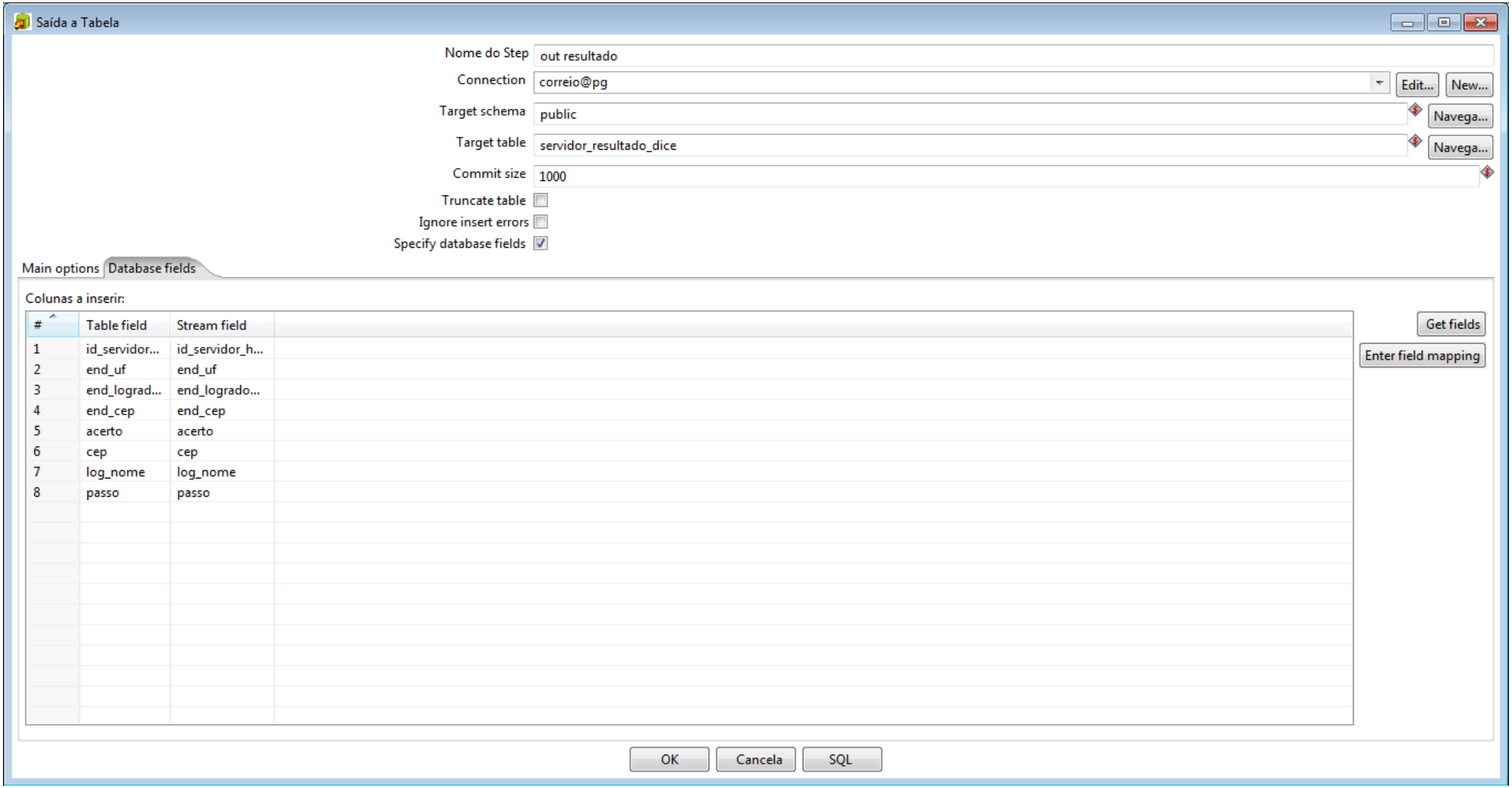


Figura 52 Out resultado - aba 2

A Figura 53 traz toda “Transformação 2 - Encontrar logradouro por meio do coeficiente de Dice” realizada por meio da ferramenta de ETL – PDI.

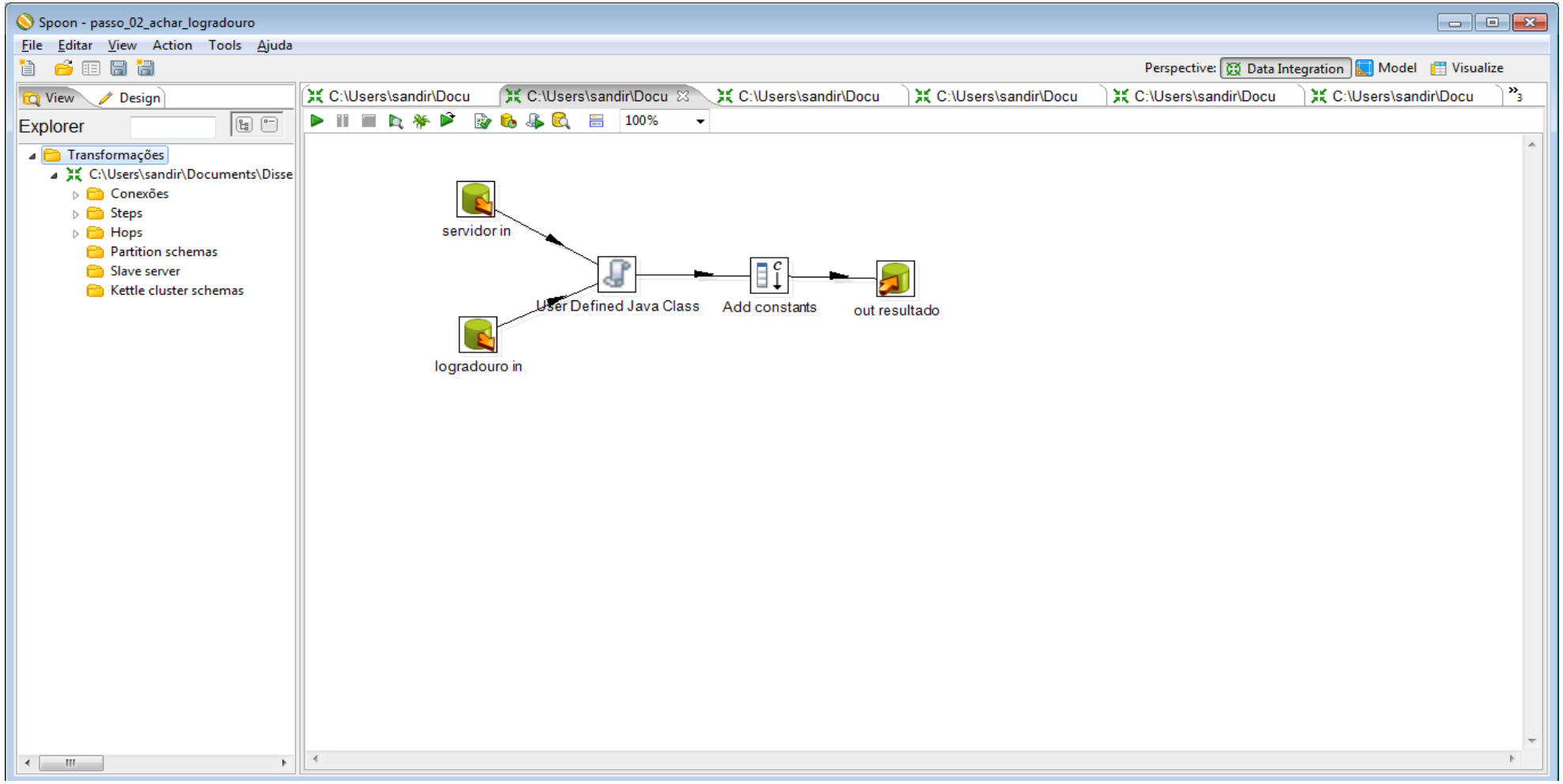


Figura 53 – Transformação 2 - Encontrar logradouro por meio do coeficiente de Dice

A transformação inicia-se com o passo do tipo *Table Input* servidor in como mostrado na Figura 26.

Ao se abrir o passo servidor in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados gerado pela extração do arquivo espelho do SIAPE na tabela siape_servidor como verificado na Figura 27.

Como visto na Figura 53, reutiliza-se a entrada de dados para a realização da "Transformação 2 - Encontrar logradouro por meio do coeficiente de Dice", o comando SQL executado na tabela siape_servidor traz os seguintes campos: id_servidor_historico(identificador único do servidor), end_uf (UF do endereço do servidor), end_logradouro em caixa alta para que possa ser comparada com a base do DNE e por fim o campo end_cep (traz o CEP cadastrado para o servidor).

O comando SQL de seleção realizado, também condiciona aos registros que não possuírem o campo end_cep com algum valor inserido e o campo end_logradouro sem nenhum valor inserido.

A transformação continua com o passo do tipo *Table Input* logradouro in como mostrado na Figura 29.

Ao se abrir o passo logradouro in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados do DNE dentro da tabela log_logradouro, como visualizado na Figura 30.

Como visto na Figura 30, reutiliza-se a entrada de dados para a realização da "Transformação 2 - Encontrar logradouro por meio do coeficiente de Dice" o comando SQL de seleção executado na tabela log_logradouro traz os seguintes campos: log_nu_sequencial, log_nome (com comando *upper* como verificado na Figura 28) e o campo cep.

A transformação continua com o passo do tipo *User Defined Java Class*, ver Figura 48.

Ao se abrir o passo "*User Defined Java Class*" Figura 54 a função apresenta o processo que foi utilizado para realizar uma chamada de importação da função da similaridade *Dice* com os parâmetros de entrada.

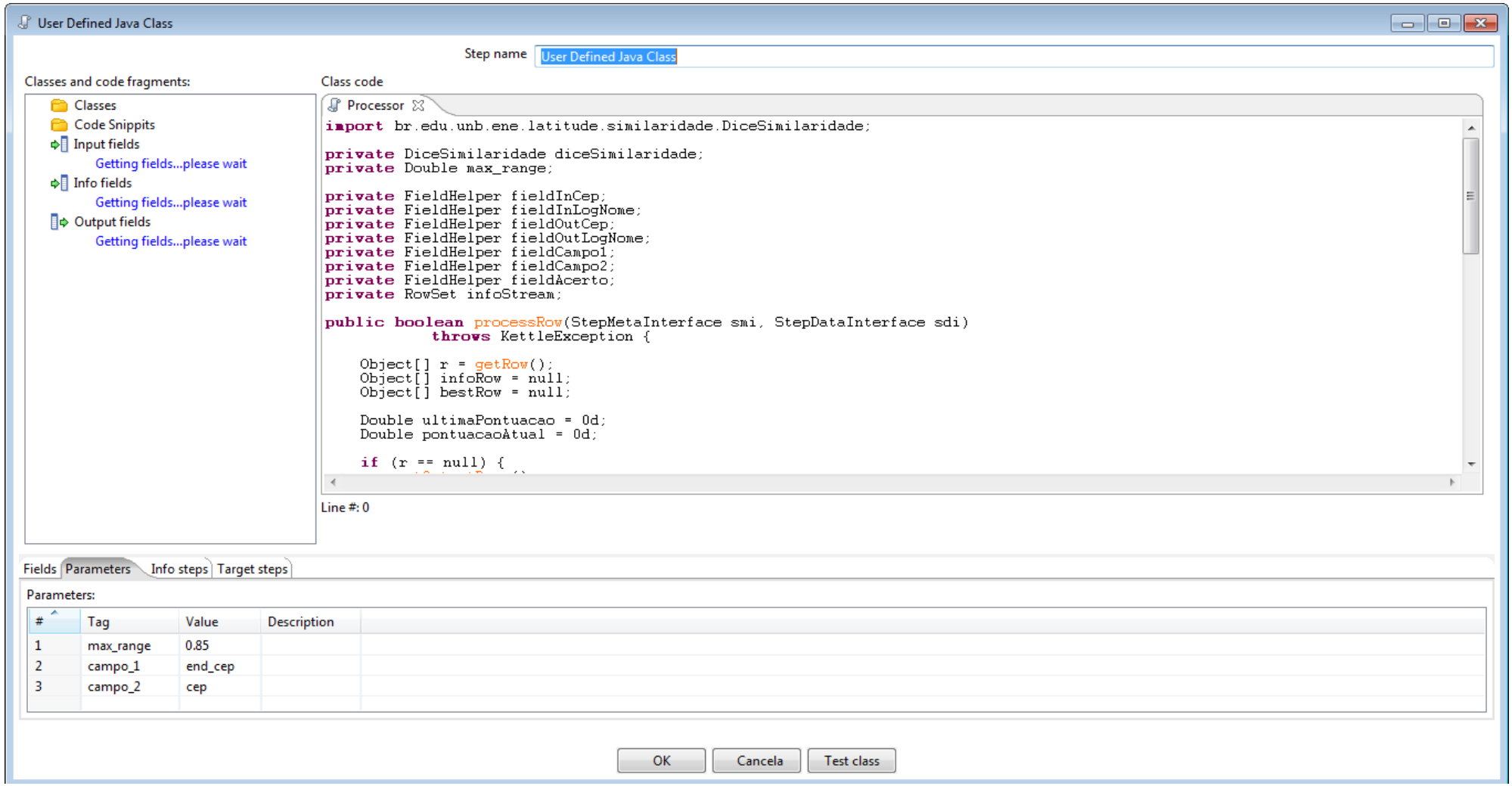


Figura 54 Java Class passo 2

A transformação continua com o passo do tipo *Add constants*, conforme Figura 33.

Ao se abrir o passo "*Add constants*" Figura 55 insere-se um campo de nome passo com valor igual a 2 para identificar dentro da tabela de saída as alterações realizadas pela Transformação 2 - Encontrar logradouro por meio do coeficiente de Dice.

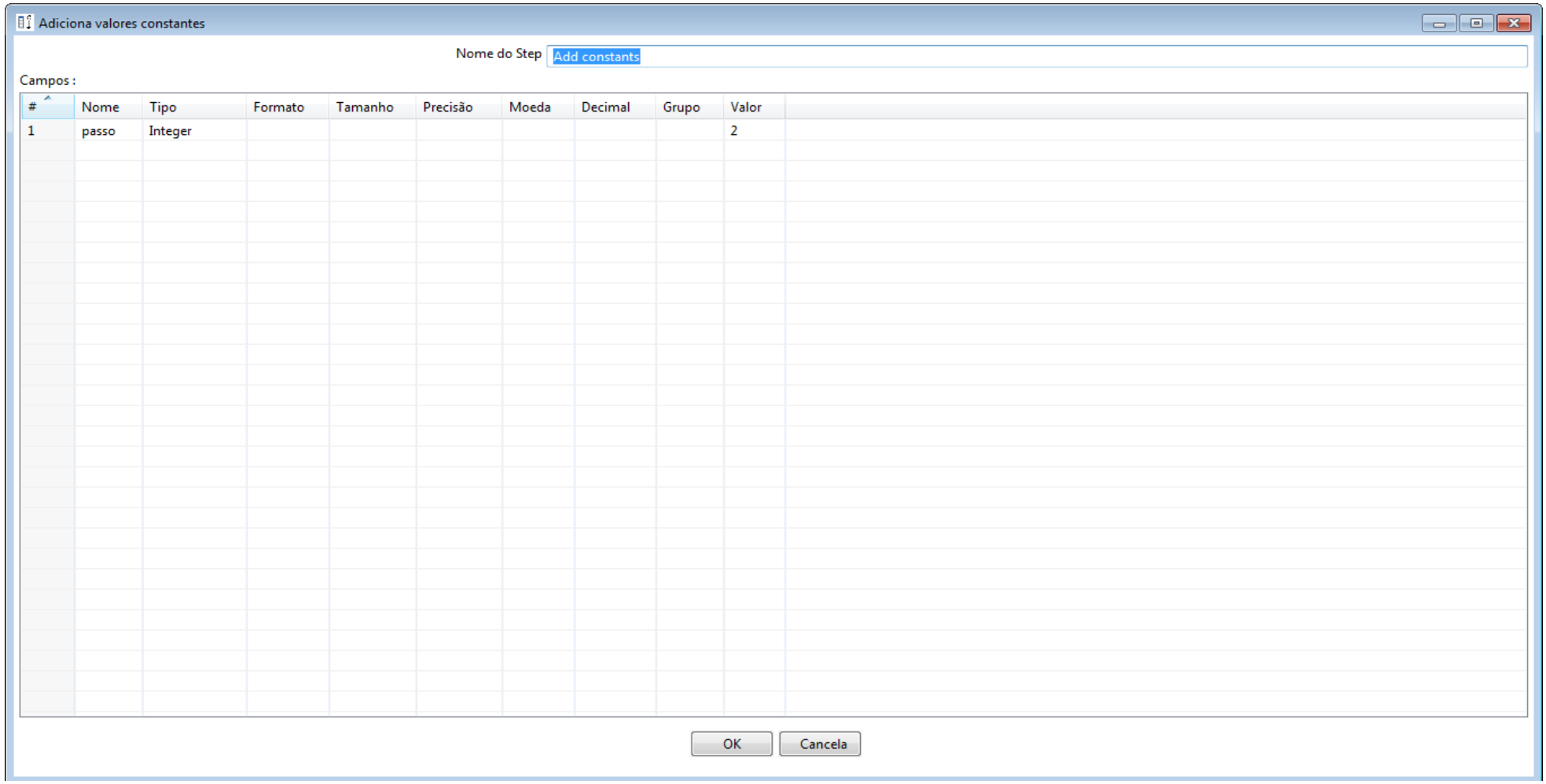


Figura 55 Add constants passo 2

A transformação continua com o passo do tipo *Table output* – out resultado, como visto na Figura 35.

Ao se abrir o passo "*Table output*", ver Figura 51 e Figura 52, inserem-se ou altera-se o nome do passo, o nome da conexão, o nome do esquema e o nome da tabela. Além disso, verifica-se o tamanho do *commit* e se a tabela deve ser truncada antes da inicialização, na segunda aba do *table* output, a aba *Database fields*, ver Figura 52, verifica-se a estrutura da tabela que será gerada, bem como, todos os seus campos correspondentes.

A Figura 56 traz toda "Transformação 3 - Completar CEP genérico por meio do coeficiente de Dice" realizada por meio da ferramenta de ETL – PDI.

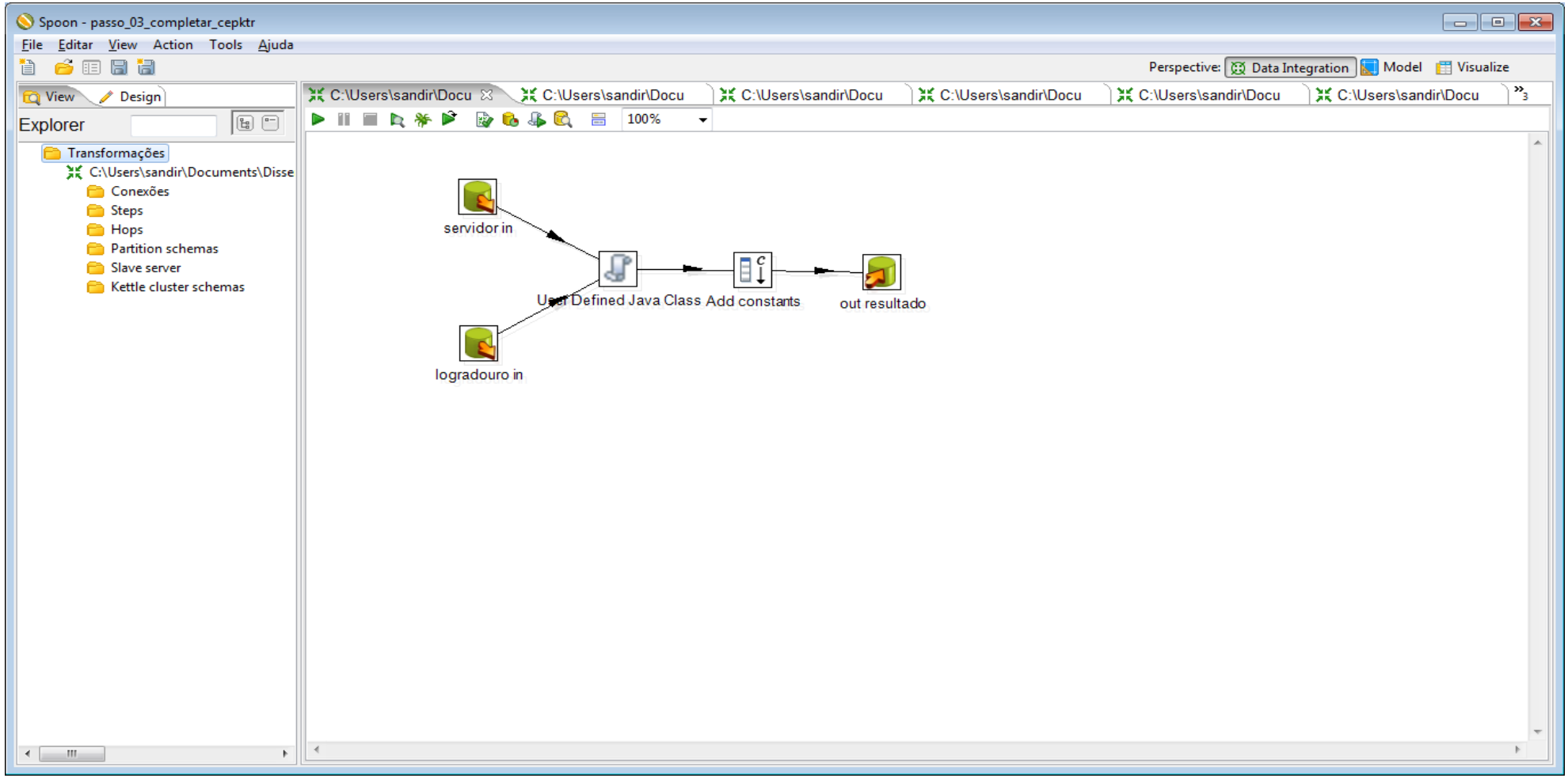


Figura 56 – Transformação 3 - Completar CEP genérico por meio do coeficiente de Dice

A transformação inicia-se com o passo do tipo Table Input servidor in como mostrado na Figura 26.

Ao se abrir o passo servidor in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados gerado pela extração do arquivo espelho do SIAPE na tabela siape_servidor, ver Figura 27.

Como visto na Figura 27, reutiliza-se a entrada de dados para a realização da "Transformação 3 - Completar CEP genérico por meio do coeficiente de Dice", o comando SQL de seleção executado na tabela siape_servidor traz os seguintes campos: id_servidor_historico(identificador único do servidor), end_uf (UF do endereço do servidor), end_logradouro em caixa alta para que possa ser comparada com a base do DNE e por fim o campo end_cep (traz o CEP cadastrado para o servidor).

O comando SQL de seleção realizado, também condiciona aos registros que não possuírem o campo end_cep e o campo end_logradouro sem nenhum valor inserido além de só trazer os valores do campo end_cep que possuírem os três últimos dígitos com valor igual a zero.

A transformação continua com o passo do tipo *Table Input* logradouro in como mostrado na Figura 29.

Ao se abrir o passo logradouro in apresenta-se o comando SQL de seleção executado na busca pelos campos específicos no banco de dados do DNE dentro da tabela log_logradouro, como visualizado na Figura 30.

Como visto na Figura 30, reutiliza-se a entrada de dados para a realização da "Transformação 3 - Completar CEP genérico por meio do coeficiente de Dice", o

comando SQL de seleção executado na tabela log_logradouro traz os seguintes campos: log_nu_sequencial, log_nome (com comando upper como verificado na Figura 28) e o campo cep.

A transformação continua com o passo do tipo *User Defined Java Class*, ver Figura 48.

Ao se abrir o passo "*User Defined Java Class*" Figura 57 a função apresenta o processo que foi utilizado para realizar uma chamada de importação da função da similaridade Dice com os parâmetros de entrada.

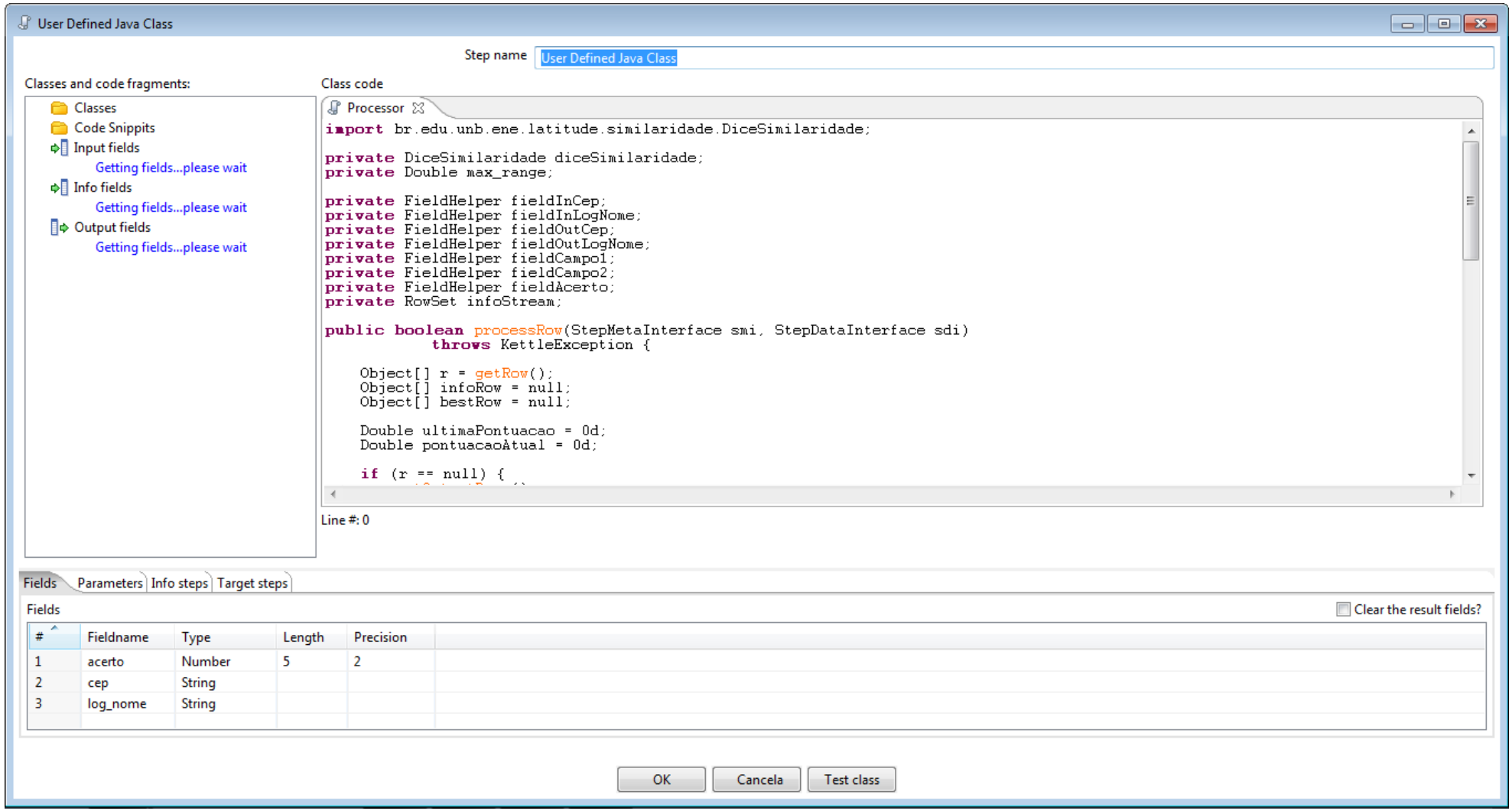


Figura 57 Java class passo 3

A transformação continua com o passo do tipo *Add constants*, conforme Figura 33.

Ao se abrir o passo "*Add constants*" Figura 58 insere-se um campo de nome passo com valor igual a 3 para identificar dentro da tabela de saída as alterações realizadas pela Transformação 3 - Completar CEP genérico por meio do coeficiente de Dice.

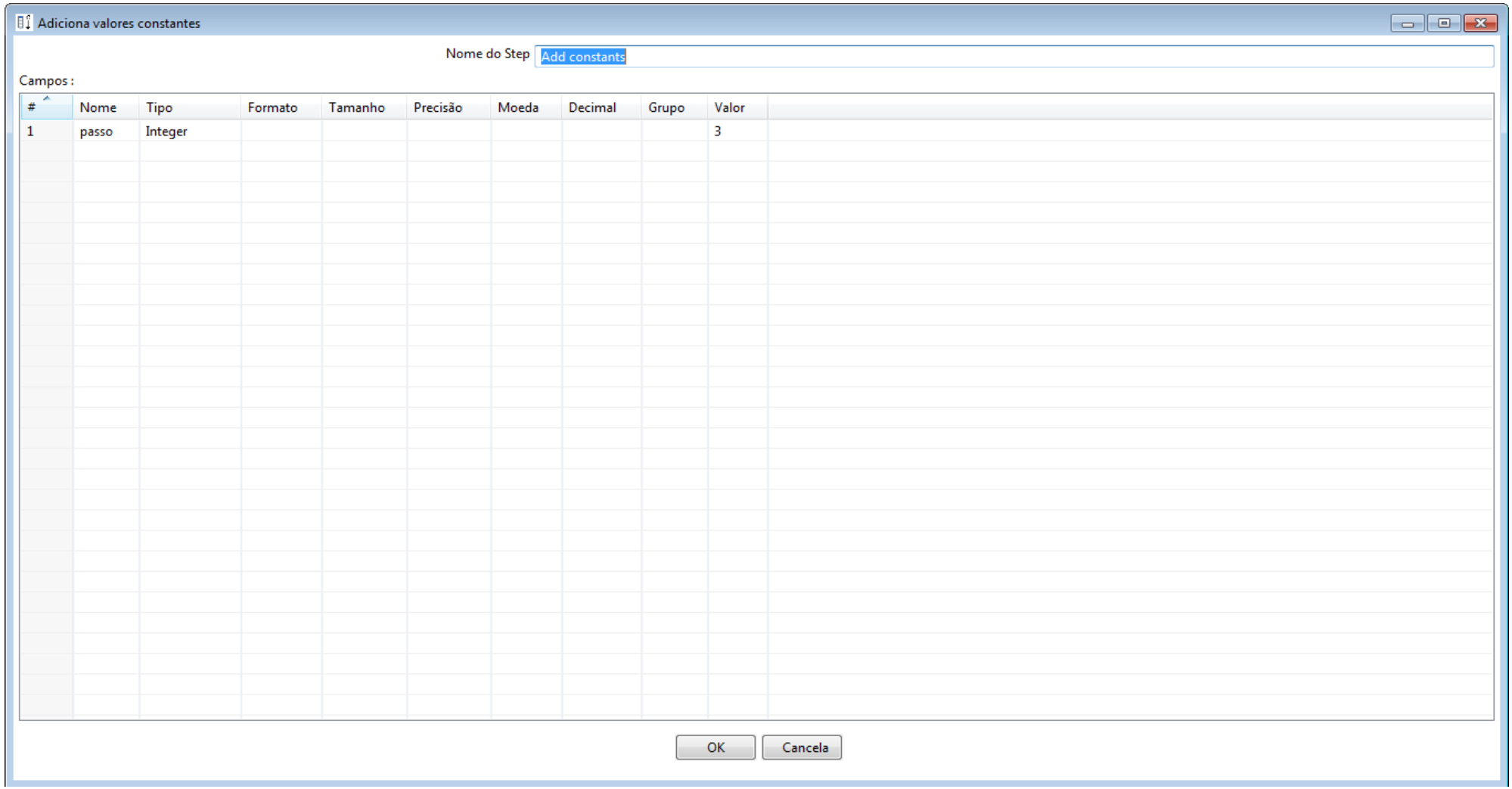


Figura 58 Add constants Dice passo 3

A transformação continua com o passo do tipo *Table output* – out resultado, como visto na Figura 35.

Ao se abrir o passo "*Table output*", ver Figura 51 e Figura 52, inserem-se ou altera-se o nome do passo, o nome da conexão, o nome do esquema e o nome da tabela. Além disso, verifica-se o tamanho do commit e se a tabela deve ser truncada antes da inicialização, na segunda aba do table output, a aba *Database fields*, ver Figura 52, verifica-se a estrutura da tabela que será gerada, bem como, todos os seus campos correspondentes.

ANEXO C – SUÍTE OPEN SOURCE PENTAHO

Diversas soluções são oferecidas no mercado para BI. O Pentaho é um software de código aberto para inteligência empresarial, desenvolvido em Java.

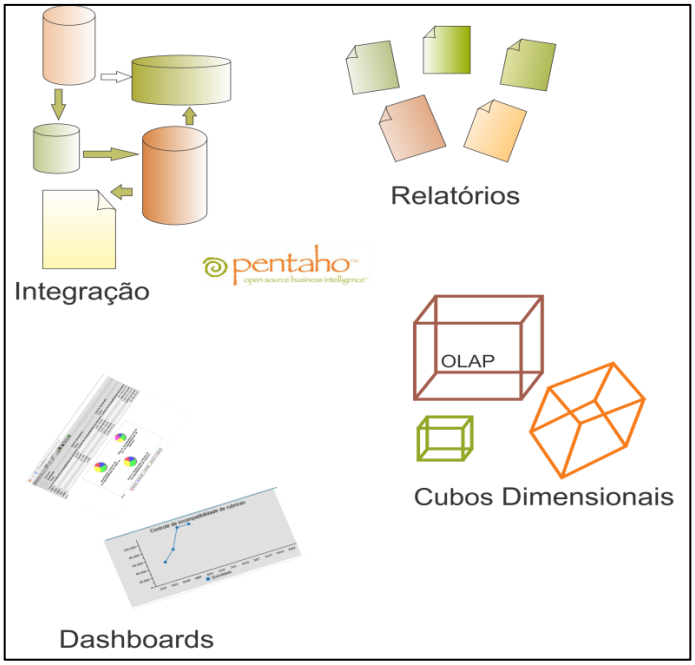


Figura 59 Ferramentas Integradas Pentaho

Conforme apresentado na Figura 59, a Suíte oferece uma completa solução para aplicação de técnicas de BI, tais como:

- *Pentaho Data Integration*(PDI): também conhecido como Kettle, é uma ferramenta de código aberto para extração, transformação e carga (ETL) de dados;
- *Pentaho Analysis Services*: também conhecido como Mondrian OLAP server, é uma ferramenta de código aberto para *On-line Analytical Processing* (OLAP);
- Pentaho *Reporting*: derivado do projeto JFreeReport, que inclui o Pentaho *Report Designer, Pentaho Reporting Engine, Pentaho Reporting SDK* e as bibliotecas comuns de comunicação compartilhadas com a plataforma Pentaho de BI. Este conjunto de ferramentas open-source de relatórios é utilizada para a criação de relatórios relacionais e analíticos de uma ampla gama de fontes de dados e tipos de saída, incluindo: PDF, Excel, HTML, Texto, Rich-Text File e XML e CSV;

- Pentaho Data Mining: derivado do projeto WEKA, incorpora uma coleção de algoritmos de aprendizado de máquina aplicados a tarefas de mineração de dados. Esses algoritmos são combinadas com tecnologias OLAP a fim de fornecer informações para análise de dados e soluções de processos, incluindo-se a criação de modelos preditivos;
- Pentaho Dashboard : inclui a apresentação de painéis e gráficos intuitivos e interativos;
- Pentaho for Apache Hadoop, também conhecido como Pentaho BI Suíte, é a versão para Hadoop, uma plataforma de software em Java de computação distribuída voltada para clusters e processamento de grandes massas de dados.

A solução Pentaho define-se a si mesma como uma plataforma de BI orientada para a solução e centrada em processos. Ou seja, não só apresenta os resultados de uma forma única e dando uma visão geral do estado da empresa, como programa os próprios processos (workflow) para a resolução de problemas detectados e apresentados.

Devido à sua estrutura em componentes, a Suíte pode ser utilizada para atender demandas que vão além do escopo das Soluções de BI mais tradicionais. Estão disponíveis componentes para a implantação de processos comandados por workflow automatizado, portais web customizáveis com suporte à *port lets* e single sign-on, entre outros.

A plataforma executa todas as suas Soluções de BI, como serviços, e por isso é possível até mesmo prover acesso a esses recursos para sistemas externos, via *web services*, por meio de um mecanismo baseado em SOAP/WSDL/UDDI incluso.

ANEXO D – ARQUIVO FONTE – DICE SIMILARIDADE

```
// INICIO ARQUIVO: DiceComparator.java

package br.edu.unb.ene.latitude.ext;

import java.util.ArrayList;

public class DiceComparator {

	/**
	 * Metodo que computa os pares de caracteres contidos em uma
string de entrada.
	 *
	 * @param str
	 * @return
	 */
	private static String[] letterPairs(String str) {
		int numPairs = str.length() - 1;
		String[] pairs = new String[numPairs];
		for (int i = 0; i < numPairs; i++) {
			pairs[i] = str.substring(i, i + 2);
		}
		return pairs;
	}

	private static ArrayList wordLetterPairs(String str) {
		ArrayList allPairs = new ArrayList();

		String[] words = str.split("\\s");
		for (int w = 0; w < words.length; w++) {
			String[] pairsInWord = letterPairs(words[w]);
			for (int p = 0; p < pairsInWord.length; p++) {
				allPairs.add(pairsInWord[p]);
			}
		}
		return allPairs;
	}

	public static double compareStrings(String str1, String str2)
{
		ArrayList pairs1 = wordLetterPairs(str1.toUpperCase());
		ArrayList pairs2 = wordLetterPairs(str2.toUpperCase());
		int intersection = 0;
		int union = pairs1.size() + pairs2.size();
		for (int i = 0; i < pairs1.size(); i++) {
			Object pair1 = pairs1.get(i);
			for (int j = 0; j < pairs2.size(); j++) {
				Object pair2 = pairs2.get(j);
				if (pair1.equals(pair2)) {
					intersection++;
					pairs2.remove(j);
					break;
				}
```

```
                }
            }
            return (2.0 * intersection) / union;
      }
}
// FIM ARQUIVO: DiceComparator.java



// INICIO ARQUIVO: DiceSimilaridade.java
package br.edu.unb.ene.latitude.similaridade;

import br.edu.unb.ene.latitude.ext.DiceComparator;

public class DiceSimilaridade {

      final static String regexSpliter = "[ ]+";


      /**
       * Cria um array de strings separando a original por espacos
e calcula o coeficiente dice para cada uma.
       * O resultado e a media dos coeficientes do array de
strings.
       *
       * @param str1
       * @param str2
       * @return
       */
      public Double compare(String str1, String str2) {

            DiceComparator diceComparator = new DiceComparator();
            Double retorno = Double.MAX_VALUE;

            String[] wrds1 = str1.split(regexSpliter);
            String[] wrds2 = str2.split(regexSpliter);

            Double sumPoints = 0d;
            Integer totalWords =
(wrds1.length>wrds2.length?wrds1.length:wrds2.length);

            String wrd1 = null;
            String wrd2 = null;

            for(int i=0;i<totalWords;i++) {
                  try {
                        wrd1 = wrds1[i];
                  } catch (ArrayIndexOutOfBoundsException ae) {
                        wrd1 = " ";
                  }

                  try {
                        wrd2 = wrds2[i];
                  } catch (ArrayIndexOutOfBoundsException ae) {
                        wrd2 = "  ";
```

```
                }

                sumPoints += diceComparator.compareStrings(wrd1,
wrd2);

           }

           retorno = sumPoints/totalWords;

           return retorno;

     }

     public static void main(String[] args) {
           DiceSimilaridade ds = new DiceSimilaridade();
           System.out.println(ds.compare("QS 305  CONJUNTO 01
LOTES", "QR 305 CONJUNTO 11"));
     }

}
// FIM ARQUIVO: DiceSimilaridade.java
```