UNIVERSIDADE TECNOLÓGICA DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

GUILHERME VINICIUS MARINASCO
TULIO MARTINS FREITAS

DATA WAREHOUSE: ESTUDO DE CASO EM UM ÓRGÃO PÚBLICO

MONOGRAFIA

CURITIBA
2015

GUILHERME VINICIUS MARINASCO
TULIO MARTINS FREITAS

DATA WAREHOUSE: ESTUDE DE CASO EM UM ÓRGÃO PÚBLICO

Monografia apresentada ao Departamento Acadêmico de Informática da Universidade Federal do Paraná como requisito parcial para a obtenção do título de “Bacharel em Sistemas de Informação”

Orientadora: Prof. Dra. Nádia Puchalski Kozievitch

CURITIBA
2015

AGRADECIMENTOS

Agradecemos a nossa orientadora Dra. Nádia Puchalski Kozievitch, por ter nos guiado e motivado no decorrer do desenvolvimento deste Trabalho de Conclusão de Curso.

À equipe do Tribunal de Contas do Paraná e a professora Dra. Maria Cláudia F. Pereira Emer pela oportunidade e apoio oferecidos por meio do convênio TCE-PR/UTFPR.

A esta universidade, seu corpo docente, direção e administração que oportunizaram a janela que hoje vislumbro um horizonte superior, eivado pela acendrada confiança no mérito e ética aqui presentes.

A todos que contribuíram direta ou indiretamente no desenvolvimento desta monografia.

DEDICATÓRIA

(Tulio)

Dedico este trabalho à minha família, pelo amor, incentivo e apoio incondicional, sem vocês nada disso seria possível. Essa vitória é não é só minha, é nossa!

(Guilherme)

Dedico este trabalho para todos que de alguma forma contribuíram para a sua realização.

RESUMO

MARINASCO, Guilherme & FREITAS, Tulio Martins. DATA WAREHOUSE: ESTUDO DE CASO EM UM ÓRGÃO PÚBLICO. Monografia – Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

A necessidade em aprimorar e agregar valor aos processos de análise de dados existentes no TCE-PR atualmente, motivou por parte dos autores a busca, estudo e pesquisa na área de soluções de data warehouse, visando assim, agregar novas soluções, melhorias, resolução de problemas e desafios de integração de dados oriundos de diferentes sistemas e bases de dados do estado do Paraná.

Suportados por uma grande quantidade de sistemas que auxiliam o gerenciamento das suas atividades rotineiras, o estágio atual dos sistemas de informação do TCE, embora já existente, não possibilitam a integração de dados entre bases de diferentes órgãos governamentais, também como a utilização de técnicas para melhoria de desempenho, tratamento de dados e controle dos problemas relacionados à carga incremental, o que inviabilizam o ambiente para a consolidação de dados gerenciais voltados para a tomada de decisão.

Com essa motivação, a presente dissertação procura prover soluções de criação de um protótipo de *data warehouse*, com o objetivo de criar relatórios específicos, fornecendo métodos eficientes para a coleta, limpeza e análise de dados. A identificação destes métodos permitiu a geração de relatórios gerencias por meio dos processos de integração dos dados.

Palavras-chave: Data warehouse, TCE-PR, sistemas de informação, armazém de dados, dados governamentais, dados abertos.

ABSTRACT

MARINASCO, Guilherme & FREITAS, Tulio Martins. DATA WAREHOUSE: CASE STUDY OF A GOVERNMENT AGENCY. Monografia – Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

The need to improve and add value to existing data analysis processes in TCE-PR, motivated the authors to search, study and research in the area of data warehouse solutions, thus aiming to add new solutions, improvements, resolution problems and data integration challenges from different systems and Parana state databases.

Supported by a large number of systems, the current staging of the TCE-PR's environment did not support the integration of data between databases from different government agencies as well as the use of techniques for performance improvement, data processing and control of problems related to incremental load, which prevented the environment for consolidating data management focused on decision making.

With this motivation, this thesis provides solutions for creating a data warehouse prototype, and providing efficient methods for collecting, cleaning and analyze the data. The identification of these methods has enabled the extraction of management reports through data integration processes.

Keywords: Data warehouse, TCE-PR, information systems, government agencies.

LISTA DE FIGURAS

LISTA DE TABELAS E QUADROS

LISTA DE ABREVIATURAS E SIGLAS

Bps: Bits por segundo.

CGI: *Common Gateway Interface* - Interface de Porta Comum.

DW: *Data Warehouse.*

IPPUC: Instituto de Pesquisa e Planejamento Urbano de Curitiba.

SQL: *Structured Query Language.*

CSV: *Comma Separated Values.*

TCU: Tribunal de Contas da União.

TCE-PR: Tribunal de Contas do Paraná.

SIAFI: Sistema Integrado de Administração Financeira.

IPARDES: Instituto Paranaense de Desenvolvimento Econômico e Social.

OLAP: *Online Analytical Processing.*

OLTP: *Online Transactional Processing.*

SIAP: Sistema de Informação de Atendimento ao Público.

SIVISA: Sistemas Operacionais da Vigilância Sanitária.

UNPAP: *United Nations Public Administration Programme*.

MD: Mineração de Dados.

SUMÁRIO

1. Introdução

A todo o momento, uma expressiva quantidade de dados são gerados dentro de corporações, envolvendo informações sobre os mais variados aspectos, tanto operacionais quanto gerenciais, armazenados em repositórios específicos. Tais repositórios, conhecidos atualmente como *data warehouse*, tornaram-se comuns na década de 90, passando a ser utilizados para obter maior eficiência no planejamento e no gerenciamento empresarial.

Armazenamento de dados, ou *data warehouse*, tem sido citado como o projeto de maior prioridade pós-milênio por mais da metade de executivos de tecnologia da informação (Sen *et al*., 2005). Porém, uma grande quantidade de ferramentas e metodologias surgiram para suportar tais aplicações, muitas vezes não atendendo essa demanda de maneira satisfatória, a fim de garantir agilidade nos processos com qualidade e consistência dos dados.

De acordo com Kimbal (1996), *data warehouse* pode ser definido como um sistema utilizado para extrair, limpar, adaptar e entregar dados históricos, com diferentes padrões, para um banco de dados dimensional com a finalidade de suportar análises baseadas em informações confiáveis e apoiar o processo de tomada de decisão. Conceitualmente, o surgimento do *data warehouse* deu-se devido à necessidade de separar os ambientes operacionais e analíticos das empresas, sendo que, uma mistura de queries analíticas e rotinas transacionais, inevitavelmente reduziria a velocidade dos sistemas, não atendendo assim a necessidade dos usuários de ambos os tipos de consulta.

Sendo assim, o processo de obtenção e extração de informações estratégicas, relativas ao contexto de tomada de decisão, é de suma importância para o sucesso de uma empresa permitindo um planejamento antecipado frente às mudanças de um mercado globalizado. Porém, tais dados geralmente encontram-se espalhados em diferentes sistemas, *hardwares* e plataformas, não havendo entre eles nenhuma forma de integração e padronização. Como solução para tal problema, diferentes empresas e indústrias, assim como agências governamentais perceberam os significantes benefícios que a implementação de um *data warehouse* poderia trazer, provendo excelentes resultados para a transformação de dados em informações consistentes de

fácil acesso, a fim de obter maior eficiência e credibilidade no gerenciamento de suas atividades (Ballard *et al*., 1998).

Analisando todos os pontos citados acima, é evidente que o *data warehouse* tem diversas aplicações que estão sendo utilizados nos mais variados campos de estudo, não estando limitado à empresas privadas, mas também variam de epidemiologia para a demografia, da ciência natural para a educação (List *et al*., 2002). É relevante citar a importância que tal assunto representa em áreas como finanças, indústrias de cartão de crédito, serviços de telecomunicação, saúde entre outros que necessitam de suporte para a tomada de decisão, envolvendo assim processos como data *mining e web mining* além de sistemas de apoio à decisão

Sistemas de apoio à decisão são os principais componentes do *data warehouse*, e são definidos como um conjunto de técnicas de TI, interativos assim como ferramentas projetadas para processamento e análise de dados para apoio dos gestores (Golfarelli *et al*., 2009). Como exemplo de fontes de dados relevantes para os sistemas de apoio à decisão que estarão presentes neste trabalho, pode-se citar informações acessíveis na base do Instituto Paranaense de Desenvolvimento Econômico e Social (IPARDES)[1], incluindo assim dados das áreas física, econômica, social, financeira, política e administrativa, disponíveis por municípios, total do Estado e para as seguintes agregações: microrregiões geográficas do IBGE[2], regiões geográficas, regiões metropolitanas e regiões administrativas do Paraná (planejamento, saúde, educação, trabalho, agricultura e comarcas/foros).

Neste contexto, observam-se projetos por parte do governo, como o SIGA[3] Brasil, que de acordo com Bastos (2009), trata-se de um sistema de informação que reúne dados do Sistema Integrado de Administração Financeira (SIAFI)[4], com objetivo de proporcionar acesso facilitado a informações por meio de um único aplicativo de tecnologia da informação. Segundo estatísticas de utilização do SIAFI, em 2009 o

[1] http://www.ipardes.gov.br - Acesso em 02.12.2015

[2] http://www.ibge.gov.br - Acesso em 02.12.2015

[3] http://www12.senado.gov.br/orcamento/sigabrasil - Acesso em 02.12.2015

[4] http://www.tesouro.fazenda.gov.br/siafi - Acesso em 02.12.2015

número de transações financeiras foi superior a um bilhão, mensurando a importância deste para a administração pública (Silva, 2010).

Devido a essa enorme quantidade de dados gerados diariamente, o Tribunal de Contas do Estado do Paraná, cuja principal função é a fiscalização da utilização do dinheiro público (Conhecendo o Tribunal, 2011), busca por meio da solução *data warehouse*, os benefícios oferecidos pela tecnologia e o apoio que este recurso pode proporcionar para garantir a seriedade na gestão dos recursos públicos. Portanto, com a produção deste trabalho de conclusão de curso, propõem-se soluções de criação de um modelo, manutenção e melhorias de um armazém de dados para que assim possa ser utilizado e replicado em outros órgãos.

1.1 Justificativa

Nos últimos anos, o *data warehouse* tornou-se um componente essencial nos sistemas de suporte a decisão, são capazes de proporcionar acesso eficiente a informações de fontes heterogêneas para fornecer suporte no planejamento e tomada de decisão nas companhias (List *et al*., 2002). Segundo Imhoff (2003), executivos esperam informações que proporcionem apoio nas decisões para conduzir suas empresas para a próxima década.

De acordo com Ballard *et al*. (2005), consolidar os dados empresariais é um passo importante para se obter o controle de uma organização, e gerenciar a partir de uma perspectiva empresarial é a chave para alcançar os seus objetivos. É a única forma de proporcionar gestão por meio da "*Visão única da empresa, ou uma versão única da verdade*", que é tão desejado, e necessário.

A necessidade em analisar e melhorar os processos existentes no *data warehouse* em desenvolvimento no Tribunal de Contas do Paraná, motivou por parte da entidade a busca e incentivo ao estudo e pesquisa na área, visando agregar novas soluções, melhorias, resolução de problemas e desafios de integração de dados oriundos de diferentes sistemas e bases de dados do estado do Paraná.

O estágio atual dos sistemas de informação do TCE, embora já existente, não possibilitava a integração de dados entre bases distintas, muito menos da utilização de

técnicas para melhoria de performance, tratamento de dados e controle dos problemas relacionados a carga incremental. A integração das bases de dados dos sistemas de informações é pré-requisito para qualquer avanço destes sistemas, que somente após integrá-las será possível uma manipulação inteligente do enorme volume disponível de dados e, consequentemente, a produção de informação relevante que contribua com as ferramentas de gestão pública (Pires, 2011).

Outro aspecto relevante para aprimoramento, refere-se ao carregamento dos dados e suas características temporais, processo esse não realizado atualmente, onde é observado a necessidade de manter informações históricas, uma vez que os usuários de sistemas de apoio a decisão usualmente estão interessados no histórico de como os dados dos provedores evoluíram ao longo do tempo.

Portanto, observa-se a necessidade de elaboração de um projeto, que contemple os componentes e requisitos de uma arquitetura de DW e suas ferramentas, considerando não apenas as etapas de desenvolvimento, mas também, a estimativa, o projeto de melhoria e um conjunto de atividades de manutenção em DW. A unificação destas informações em um único ambiente de forma integrada e padronizada tornará possível o aprimoramento de sistemas de suporte a decisão, indicadores econômicos, análise de tendências e demais informações de interesse público. Busca-se assim reduzir a quantidade de erros nos processos atuais, garantir a consistência dos dados e fornecer soluções para os problemas de desempenho, extração, transformação e carga, assim como facilitar, a comunicação e integração entre órgãos públicos.

1.2 Objetivo Geral

Realizar um estudo de caso fundamentado na criação de um *data warehouse* para um órgão público para geração de relatórios específicos, listando os principais desafios, soluções de integração e possíveis melhorias.

1.3 Objetivos Específicos

1. Revisar a literária sobre: (i) criação de *data warehouse* e aplicações e (ii) dados governamentais;
2. Revisar os padrões atuais, visando a qualidade e estrutura dos dados;
3. Obter e integrar os dados de fontes heterogêneas;
4. Apresentar os processos, metodologias e técnicas utilizadas no desenvolvimento do protótipo de um *data warehouse* que constitui a parte prática desta dissertação;
5. Realizar estudo de caso em um contexto específico, buscando explorar problemas de integração entre órgãos públicos por meio da:
    - Verificação de inconsistências;
    - Verificação de padrões;
    - Utilização de técnicas de *data warehouse;*
    - Verificação e utilização da base de dados e
    - Implementação de um protótipo de *data warehouse* para geração de relatórios específicos.

1.4 Estrutura/Organização

Este documento de Trabalho de Conclusão de Curso está organizado em 7 capítulos incluindo esta introdução, e 3 apêndices sendo:

- **Capítulo 2 – Trabalhos Relacionados:** Nesta seção é apresentado o referencial teórico do trabalho incluindo conceitos e análises necessárias para o desenvolvimento da proposta, e também trabalhos correlatos significantes para a área de estudo.
- **Capítulo 3 – Metodologia:** Neste tópico é apresentada a estratégia do desenvolvimento do trabalho.
- **Capítulo 4 – Recursos de Hardware e Software:** Nesta etapa, apresenta-se toda a estrutura de hardware e software necessários para o desenvolvimento do trabalho além da viabilidade do estudo de caso.
- **Capítulo 5 – Implementação:** Esta seção é composta pela parte prática proposta na metodologia deste trabalho.

- **Capítulo 6 – Conclusões:** Nesta seção, apresentam-se as conclusões geradas a partir da realização do trabalho e os trabalhos futuros.
- **Capítulo 7 – Referências:** Por fim, apresentam-se as referências utilizadas como base teórica para o desenvolvimento do trabalho.
- **Apêndice A –** Apresenta os scripts utilizados no processo de limpeza dos dados.
- **Apêndice B –** Apresenta os scripts utilizados na criação da *staging area*.
- **Apêndice C –** Apresenta os scripts utilizados para criação das tabelas do *data warehouse*.

2. Levantamento Bibliográfico e Estado da Arte

Neste capítulo apresentam-se conceitos de *data warehouse*, governo eletrônico, e considerações a respeito.

2.1 *Data Warehouse*

Com o crescimento explosivo do volume de dados armazenados atualmente, novas técnicas e serviços passaram a ser utilizados com a adoção de aplicações baseadas em *data warehouse*. Além disso, o número de usuários vem crescendo cada vez mais, dia após dia. Consequentemente, o processo de análise de dados tornou-se um processo trabalhoso, exigindo cada vez mais profissionais qualificados e ferramentas específicas para apoiarem tais atividades.

Ademais, geralmente, os dados encontram-se em sistemas heterogêneos, ou seja, espalhados por sistemas diferentes, sem qualquer forma de integração, sem qualidade e indisponíveis para os gerentes e altos executivos que são os tomadores de decisões estratégicas das organizações (Mussi *et al., 2004*). Tendo em vista essa dificuldade de analisar grandes volumes de dados, percebe-se a necessidade de novas estruturas que suportem, de forma otimizada, os conceitos de multidimensionalidade de dados e navegabilidade hierárquica facilitada.

Outro fator que contribuiu para a importância do *data warehouse* foi a necessidade de produzir informações consistentes e confiáveis, baseado em dados históricos, uma vez que dados relativos a um grande espectro de tempo (5 a 10 anos) encontram-se disponíveis (Ciferri, 2002).

Neste sentido, para suprir tais deficiências surgiu o *data warehouse*, que além de constituir um conjunto de arquiteturas e/ou sistemas de informação orientados a assunto que existem em plataformas segregadas do ambiente transacional, é uma excelente alternativa ao enfoque tradicional para integração e acesso de dados a fontes de informações heterogêneas. Sendo assim, com a utilização de tal tecnologia, torna-se possível manipular uma grande quantidade de dados, principalmente históricos, além realizar a integração de diferentes sistemas e uma única e consistente base de

dados que permitirá análises e decisões complexas de negócio (Mussi *et al., 2004*).

Disto isto, pode-se listar algumas limitações que são resolvidas ao se implementar um *data warehouse:*

- ➔ Integração: Dispersão de sistemas operacionais e base de dados;
- ➔ Credibilidade: Discrepância entre as informações contidas no BD;
- ➔ Desempenho: Duração das transações;
- ➔ Histórico: Alterações constantes nos dados;
- ➔ Redundância: Mesmos dados em bases diferentes;

2.1.1 Conceitos

A conceituação do termo *data warehouse* foi apresentada por Bill Inmon na década de 1980, que define *data warehouse* como uma base de dados orientada a assunto, integrada, não volátil e temporal, de suporte a decisões gerenciais.

De outra maneira, Ciferri (2002), define *data warehouse* como um banco de dados voltado para o suporte aos processos de gerência e tomada de decisão, e tem como principais objetivos prover eficiência e flexibilidade na obtenção de informações estratégicas e manter os dados sobre o negócio com alta qualidade.

Já Ballard *et al*. (2005), trata *de data warehouse* como implementação de processos, ferramentas e facilidades para gerenciar e fornecer informações completas, oportuna, precisa e compreensível para a tomada de decisão. Ele inclui todas as atividades que tornam possível para uma organização para criar, gerenciar e manter um *data warehouse.*

Por fim, Barquim *et al*. (1997) define *data warehouse* como um único repositório composto por dados históricos, extraídos de bases transacionais e/ou dados externos, sendo integrados e possibilitando assim a análise massiva de informações, de forma a permitir melhores tomadas de decisões e a descoberta de conhecimento, sem impactar no desempenho dos bancos de dados transacional.

Levando em consideração a última definição, podemos conceituar os ambientes que alimentam o data *warehouse*, como *Online Transactional Processing (OLTP)* e

Online Analytical Processing (OLAP) (Inmon, 1996). Sistemas OLTP são bases de dados que sofrem atualizações constantes, utilizados para suportar transações operacionais diárias (Irtishad *et al*., 2004). Normalmente, processam transações pequenas e isoladas, utilizadas para leitura ou escrita de dados, podendo assim sustentar grandes volumes de requisições simultâneas.

Porém, tais sistemas não são adequados para suportar consultas de tomada de decisão, com a finalidade de responder perguntas à nível gerencial. Para tal finalidade, processos de análise de dados (Agregação, *drilldown* e *slicing/dicing, etc*) são melhor suportados por sistemas OLAP (List *et al*., 2002). Entre outras características dos sistemas OLAP, podem ser destacadas que são utilizados somente para leitura, possuindo uma grande quantidade de dados históricos, acessados por meio de queries complexas, tendo uma característica multidimensional.

2.1.2 Características

Distribuir dados em uma base de dados centralizada levando-se em consideração as características intrínsecas de aplicações de *data warehousing* e as necessidades dos usuários típicos de sistemas de suporte à decisão representa uma área de pesquisa muito importante a ser explorada (Wu *et al*., 1997).

Kimbal (1996) definiu características que devem faser parte de um *data warehouse,* sendo elas:

1. O *data warehouse* deve tornar as informações de uma organização de fácil acesso.

As ferramentas utilizadas devem ser simples de se manipular, de modo que as aplicações possam acessar o armazém de dados e compartilhar as informações facilmente para os mais variados tipos de usuários.

2. O *data warehouse* deve apresentar a informação de uma organização de forma consistente.

Os dados devem ser colocados no *data warehouse* com um formato consistente, a fim de evitar conflitos de nomes e medidas. Esse processo é alcançado quando os dados são trazidos de diferentes bases, para um repositório central.

3. O *data warehouse* deve ser flexível e adaptável à mudanças.

Flexibilidade e adaptação estão relacionadas com as rápidas mudanças que ocorrem diariamente no ambiente operacional. Portanto, o *data warehouse* deve estar apto a aderir novas tecnologias, assim como permitir a mudança incremental de dados, utilizando múltiplas base de dados e sistemas operacionais.

4. O *data warehouse* deve servir de base para as tomada de decisão dentro de uma organização.

De acordo com Ciferri (2002) a obtenção de informações estratégicas, relativas ao contexto de tomada de decisão, é de suma importância para o sucesso de uma empresa, permitindo de tal maneira à empresa um planejamento rápido frente às mudanças.

Dito isto, Kimbal (1996) caracteriza os dados contidos no *data warehouse* em: integrado, não-volátil, variante no tempo, orientados ao assunto e nível de granularidade. Estas características são apresentadas nas próximas seções.

2.1.2.1 Integrado

O processo de integração de dados contidos em diferentes sistemas é a base para a qualidade de informações fornecidas pelo *data warehouse*. Esses dados, na maioria das vezes, são provenientes de bases heterogêneas, trazendo informações relacionadas aos processos operacionais. Deste modo, Ciferri (2002) chama a atenção para as possíveis inconsistências que podem ocorrer no processo de extração, devido a diferenças semânticas nos formatos dos dados.

Sendo assim, é importante que dados relacionados a nomes, unidades métricas e unidades em geral sejam transformados em um mesmo padrão, até atingirem um estado uniforme. Por exemplo, um sistema pode reconhecer pessoas do sexo masculino pela letra alfabética "M" (Masculino), enquanto outra base pode codificar o mesmo sexo como "H" (Homem).

Portanto, no processo de integração, para minimizar problemas de integridade e aprimorar a qualidade dos dados é necessário haver controle de qualidade, atentando-se assim com dados que apresentam diferentes versões do mesmo assunto

no mesmo banco de dados. Problemas de integridade podem levar à elaboração de relatórios inconsistentes, assim como afetar a confiabilidade de tais sistemas.

Na figura a seguir, Rudra *et al*. (1999) apresentam uma relação negativa entre a consistência e a redundância dos dados. Analisando a Figura 1, percebe-se que quanto menor a redundância, ou melhor, quanto menor à repetição não necessária dos dados contidos na base, maior é a consistência dos dados.

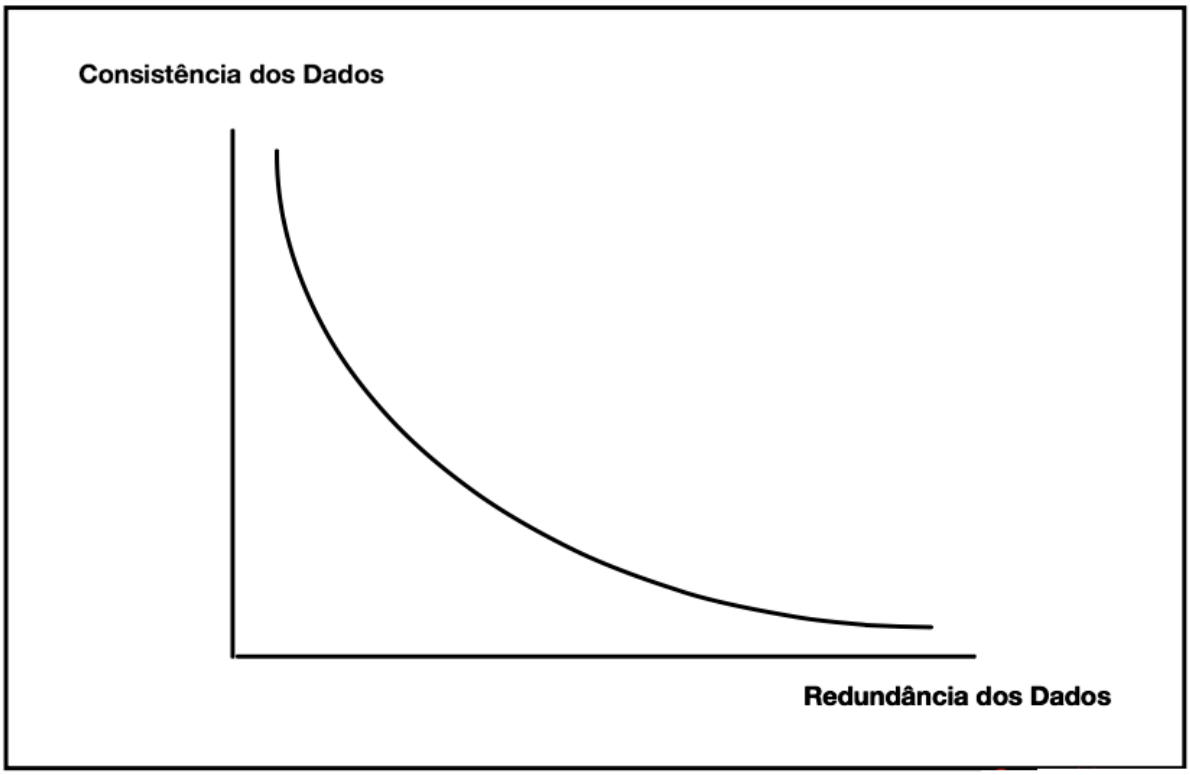


Figura 1: Relação consistência x redundância. Fonte: Rudra *et al*., (1999).

Por outro lado, existe uma relação direta entre a consistência e a integridade dos dados, como representado na Figura 2. Sendo assim, eliminando a redundância de dados é possível atingir altos níveis de consistência de dados e, consequentemente, a integridade dos dados de maneira satisfatória.

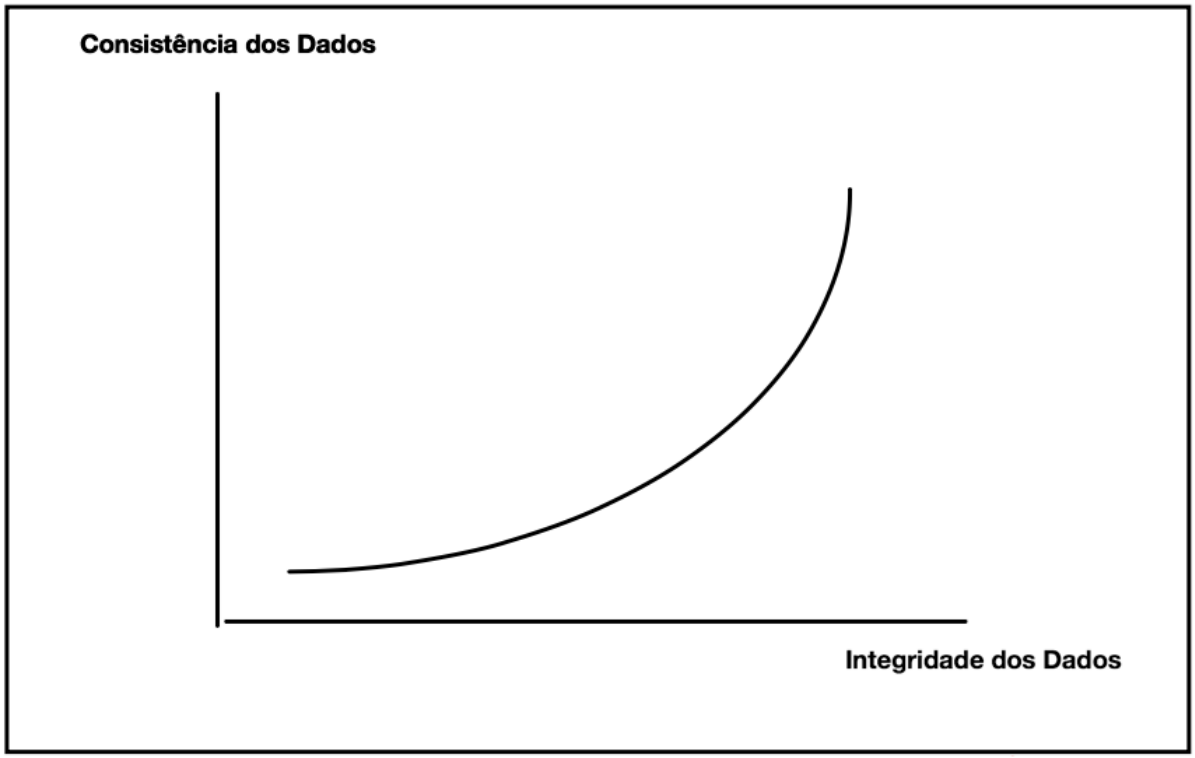


Figura 2: Relação consistência x integridade. Fonte: Rudra et al., (1999).

2.1.2.2 Não-Volátil

O conceito de não volatilidade está ligado diretamente ao fato de que o *data warehouse* permite apenas a carga inicial e consultas a dados, permanecendo assim, estável por longos períodos de tempo (Rainardi, 2008).

O processo de carga permite que novos dados sejam inseridos no sistema, que são carregados em blocos depois de terem sido integrados e transformados. Já as consultas, são caracterizadas por serem somente para leitura *(read-only),* não havendo assim consultas comuns nos ambientes operacionais, como: Incluir, excluir, alterar e etc.

2.1.2.3 Variante no Tempo

A definição de variante no tempo refere-se ao fato dos dados de um *data warehouse* serem históricos, envolvendo assim um momento específico, como por exemplo, um período de 5 anos. Portanto, difere-se do ambiente operacional, onde os dados são válidos somente para o momento do acesso. Golfarelli *et al*. (2009) faz uma

comparação visual, relacionando dados operacionais como fotografias tiradas em um certo intervalo de tempo. A sequência dessas fotografias seriam armazenadas no *data warehouse* e, os resultados poderiam ser mostrados como um filme, revelando assim a situação da empresa desde sua fundação até o momento atual.

Consequentemente, o volume de dados presente no *data warehouse* se torna muito superior ao volume presente no sistema operacional, o que poderia gerar algum problema de complexidade na administração de tal ambiente. Por fim, é importante citar que cada entrada no *data warehouse* possua um componente de tempo associado ao mesmo (Ciferri, 2002).

2.1.2.4 Orientados ao Assunto

O *data warehouse* armazena informações sobre os dados corporativos, sendo estes, específicos e importantes aos temas de negócio de maior interesse da corporação. Alguns exemplos de tema podem ser citados, como: produto, vendas, clientes e etc.

2.1.2.5 Granularidade

Conceito fundamental em projetos de *data warehouse*, Inmon (1997) destaca que a definição do nível de granularidade é um dos maiores desafios para o desenvolvedor, sendo uma etapa que permeia toda a arquitetura que envolve o ambiente de *data warehouse*, quando é apropriadamente definida, os demais aspectos do projeto e implementação não se complicam.

Enfatizando tal importância, Ciferri (2002) apresenta granularidade como o nível de detalhe em que as informações são armazenadas, sendo determinante no volume de dados alocados no *data warehouse* e no tipo de consultas que podem ser respondidas pelo sistema. Neste contexto, Inmon (1997) mostra em uma balança, conforme Figura 3, a relação entre o nível de detalhes, o volume de dados e o nível de manipulação proporcionado.

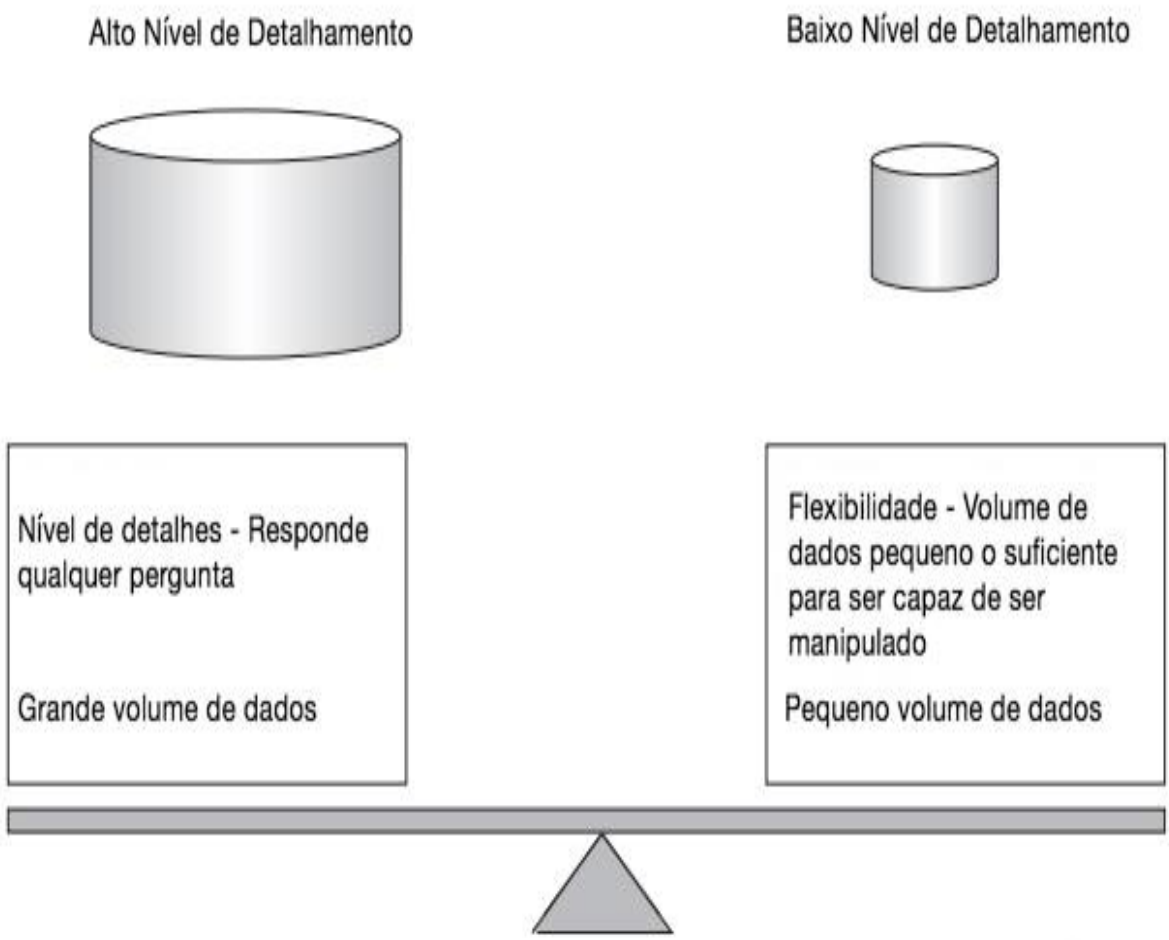


Figura 3: Relação nível de granularidade. Fonte: Inmon (1997).

Nesta relação de nível de detalhes e granularidade, Inmon (1997) explica que quanto maior o nível de detalhes de uma informação, menor é seu nível de granularidade, e exemplifica que um conjunto de transações de baixo nível de granularidade quando consolidadas podem obter um alto nível de granularidade.

Entretanto, a escolha do nível de granularidade é uma tarefa conjunta com a necessidade do negócio, conforme relaciona Ballard *et al.* (2005), os fatores apresentados na Tabela 1, possuem impacto significante sobre a determinação do nível de granularidade.

Fator	Descrição
Necessidade atual do negócio	A necessidade do negócio é o fator determinante para o nível de granularidade, espera-se no mínimo responder todas as perguntas da área de negócios no âmbito do *data warehouse*.
Necessidades futuras do negócio	A implementação do *data warehouse* deve considerar necessidades futuras do negócio, proporcionando nível de granularidade adequado para perspectivas futuras.
Necessidades adicionais do negócio	Obter informações específicas sobre o segmento do negócio, pode proporcionar maior assertividade para determinar o nível de granularidade de acordo com a área de negócio.
Necessidade de mineração de dados	Solicitações de mineração de dados exigem detalhes significativos e diretamente relacionados com o nível de granularidade.
Necessidade de dados derivados	O nível de granularidade escolhido deve acomodar o armazenamento de todos os elementos utilizados para deduzir os outros elementos de dados, exceto em casos com problemas de desempenho e custo elevado.
Granularidade dos sistemas	O nível de granularidade dos sistemas de origem, especialmente quando trata-se de sistemas com informações em níveis de detalhes diferentes, devem ser considerados para determinar o nível de granularidade do *data warehouse*.
Desempenho na aquisição de dados	O nível de granularidade pode proporcionar um impacto significativo sobre o desempenho de aquisição de dados, operações individuais podem ser necessárias no processo de extração de dados dos sistemas de origem, afetando o processo de carga do *data warehouse.*
Custo de armazenamento	O nível de granularidade possui impacto significativo no custo de armazenamento, um alto nível de detalhes implica em um grande volume de dados.
Administração	A inclusão de detalhes adicionais no *data warehouse* impacta diretamente na sua administração, como por exemplo rotinas de *back-up*.

Tabela 1: Fatores determinantes na escolha da granularidade. Fonte: Ballard *et al.*, (2005).

2.1.3 Arquitetura

No cenário de sistemas de informação, a arquitetura é fundamental para o planejamento e melhor comunicação na construção do projeto, proporcionando maior flexibilidade, produtividade e facilidade de aprendizado ao sistema (Kimball, 1996).

Conforme Ballard *et al.* (2005), a estrutura do *data warehouse* é modelada de acordo com a necessidade do negócio, e apresenta na Figura 4, um modelo genérico de repositório de dados para suporte a decisão.

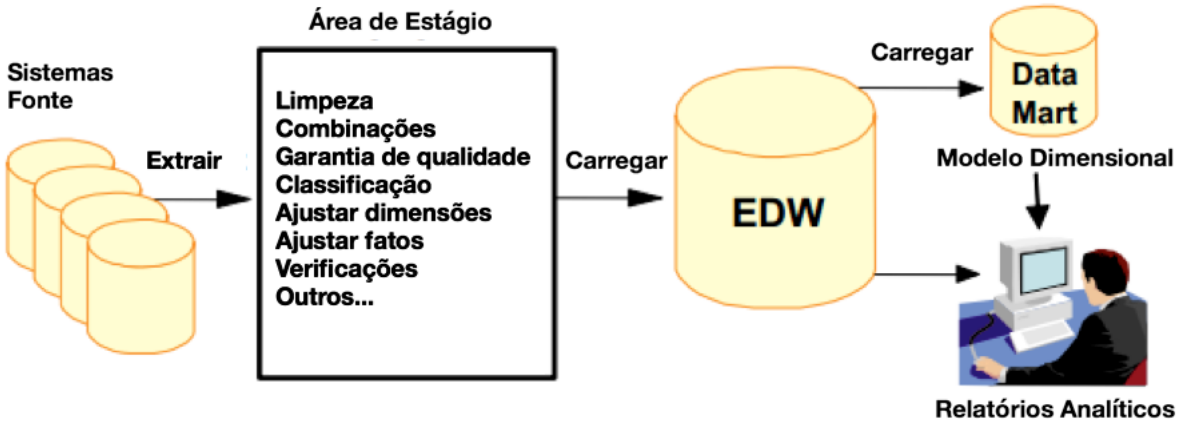


Figura 4: Modelo genérico de repositório de dados. Fonte: Ballard *et al.*, (2005).

Tratando-se de modelos genéricos, Ciferri (2002) também apresenta um modelo de arquitetura convencional de um ambiente de *data warehousing* conforme Figura 5, utilizada para criar, manter e consultar um *data warehouse*.

Seguindo no modelo de arquitetura apresentado na Figura 5, Ciferri (2002) destaca a importância do componente de integração e manutenção, o qual é responsável por uma série de atividades de preparação dos dados provenientes dos provedores de informação. Este processo será discutido com maiores detalhes no próximo tópico.

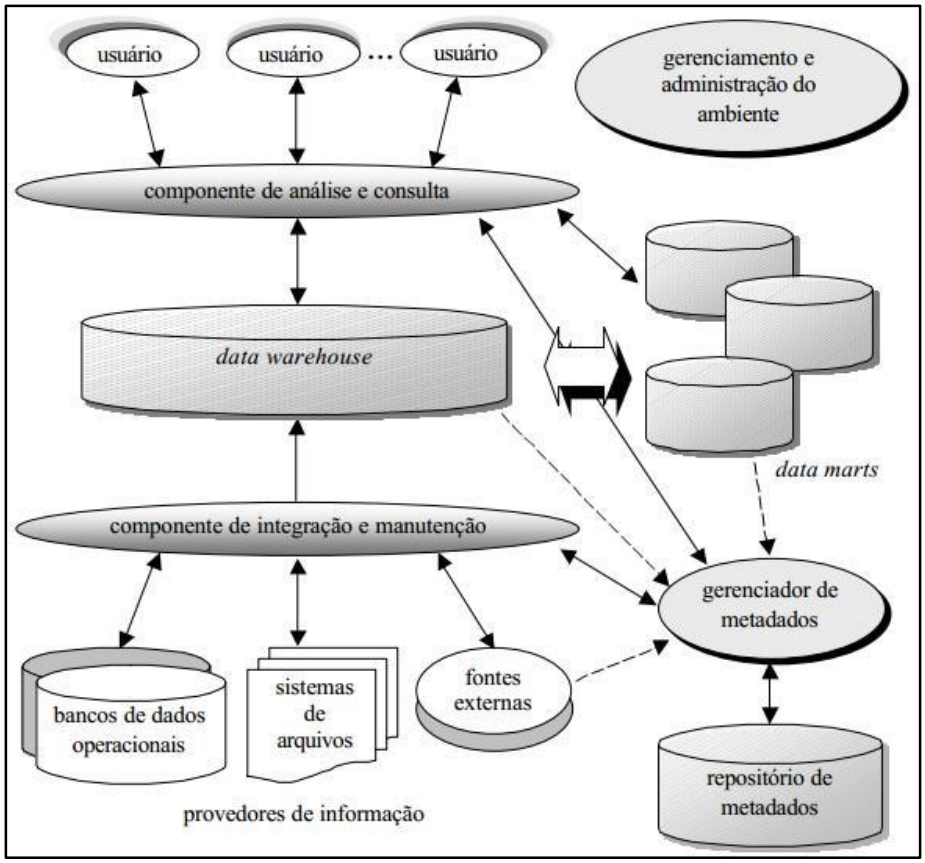


Figura 5: Arquitetura típica de um ambiente de *data warehousing*. Fonte: Ciferri (2002).

2.1.4 Componentes de um *data warehouse*

As informações inseridas no *data warehouse*, devem ser limpas e apresentar boa qualidade, para garantir estes aspectos, torna-se necessário a realização de diversas validações antes que o dado seja inserido no *data warehouse*. Tais validações visam garantir, por exemplo, a correspondência entre informações, a presença de valores inválidos ou a ausência destes (Rainardi, 2008).

Neste contexto, Ciferri (2002) intensifica a importância dos processos de extração, tradução, limpeza, integração e do armazenamento de dados oriundos dos provedores de informação na inserção no *data warehouse*, sendo estes

imprescindíveis ao bom funcionamento do ambiente. A combinação dos componentes resulta na criação de um ambiente de *data warehouse*, sendo que cada componente tem sua função específica*.*

Deste modo, Kimbal (1996) apresenta um modelo baseado em sistemas operacionais (origem dos dados); área de estágio; apresentação de dados e ferramentas de acesso aos dados. Já Santos e Gutierrez (2008) dividem o *data warehouse* em quatro elementos: dados operacionais; processo de carga (ferramentas ETL); informações analíticas (ferramentas OLAP) e metadados. Na Figura 6, apresenta-se os conceitos propostos por Kimbal (1996).

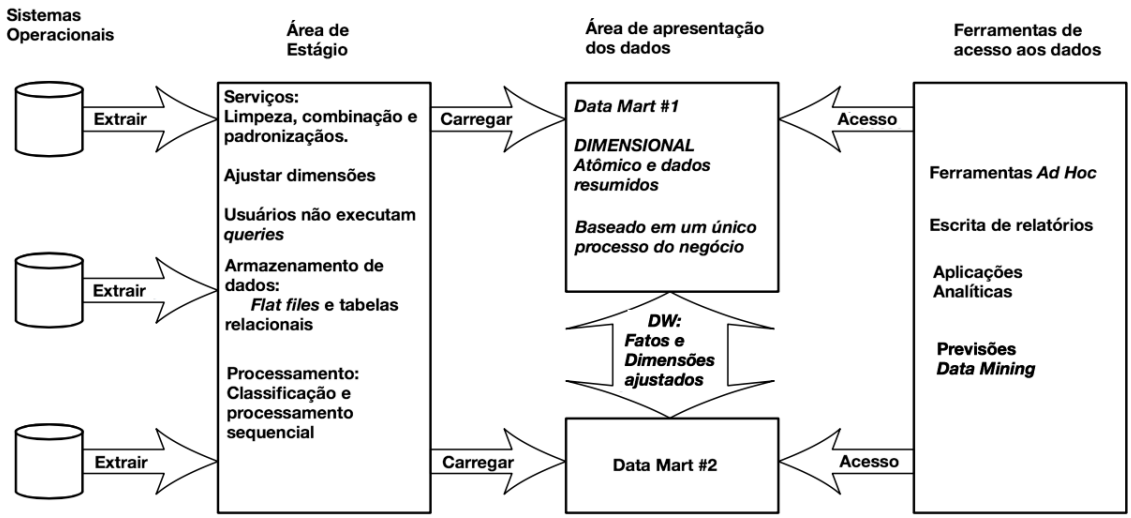


Figura 6: Componentes do DW. Fonte: Kimbal (1996).

2.1.4.1 Sistemas Operacionais

De acordo com Kimbal (1996), o *data warehouse* é composto por um conjunto de sistemas operacionais que realizam captura e armazenamento das transações diárias. Normalmente, esses sistemas são preparados para suportar um alto volume de processamento, sendo necessário alto desempenho e disponibilidade dos dados. Portanto, tais sistemas servem como base para o *data warehouse*, sendo que os dados

que serão carregados no *data warehouse*, podem estar armazenados em múltiplos sistemas operacionais.

2.1.4.2 *Staging Area* e processos ETL

A área de estágio engloba os processos de extração, transformação e integração dos dados (ETL) provindos de sistemas operacionais, para que assim, populem o *data warehouse* de maneira integrada e consistente. Em comparação com as demais fases, Golfarelli *et al*. (2009) define a fase ETL como a mais complexa, cara e demorada, representando assim, um elevado nível de importância, pois tais fases implicarão quais dados serão coletados, a maneira como serão tratados e finalmente, como serão disponibilizados ao usuário final.

Sendo assim, pode-se dizer que a área de estágio realiza uma ligação entre os sistemas transacionais e o *data warehouse*, processando, migrando e transformando dados por meio dos processos ETL, que serão definidos a seguir:

- **Extração dos Dados – *Extraction***

A extração de dados é a primeira fase do processo ETL, servindo de apoio as etapas seguintes. Neste momento, os dados são retirados dos sistemas operacionais por meio de operações de leitura e compreensão dos dados para posteriormente serem tratados na área de estágio.

Quando realizado pela primeira vez, ou seja, para popular o *data warehouse* é realizado um processo chamado de *static extraction*, ou *full extraction*, onde os dados são capturados em certo período de tempo, como uma “fotografia”. Ou seja, os dados são extraídos completamente dos sistemas de origem. Por outro lado, existe o processo de extração incremental, que captura somente as mudanças realizadas desde a última extração.

- **Transformação dos Dados – *Tranformation***

Nesta etapa busca-se a garantia de que um conjunto de dados está no formato correto e tenham um padrão de qualidade. É nesse momento que os dados inválidos e desatualizados sofrem um processo de limpeza, por meio de ferramentas específicas, para depois serem colocados no *data warehouse*. Ou seja, ocorre o processo de

identificação e correção dos dados “sujos”. Além disto, os dados são convertidos do formato do sistema fonte para o formato padrão do *data warehouse*.

Por se tratar de um ambiente voltado para a tomada de decisão, tais informações devem ser íntegras e confiáveis, caso contrário poderiam resultar em efeitos negativos e cenários incorretos, gerando um prejuízo para a corporação. Em virtude de tal importância, a limpeza dos dados deve ser realizada durante todos os demais processos ETL, e não somente como uma atividade separada (Ciferri, 2002). Sendo assim, Golfarelli *et al*. (2009) citam uma lista de erros e inconsistências mais frequentes, que devem ser tratados a fim de garantir a qualidade e integridade dos dados:

- **Dados duplicados** – Quando o *data warehouse* é alimentado por múltiplas bases de dados, um erro muito comum é a duplicação de registros.
- **Valores inconsistentes que estão logicamente associados** – Exemplo: CEP e Endereço.
- **Dados incompletos** – Exemplo: Em uma determinada tabela de empregados, o atributo “NomeDoCargo” poderia estar sem preenchimento.
- **Uso inapropriado dos campos** – Exemplo: No lugar do CPF, o usuário preencheria com um número de telefone.
- **Valores impossíveis ou errados**: Exemplo: 35/15/2015 (Data inválida).
- **Valores inconsistentes devido aos diferentes formatos** – O mesmo valor é salvo em formatos diferentes. Por exemplo: Paraná e PR.
- **Valores inconsistentes devido ao erro de digitação** – O usuário pode digitar o mesmo valor, porém, devido a erros de digitação, são tratados como valores distintos. Exemplo: Rua Jose de Sousa / R. Jose de Souza.

Devido ao fato de cada base estruturar seus dados diferentemente, os dois últimos itens são muito comuns em ambientes de *data warehouse* sustentados por mais de um banco de dados.

Faria (2006) apresenta em seu trabalho, uma solução para a integração entre sistemas operacionais da Vigilância Sanitária (SIVISA) e o Sistema de Informação de Atendimento ao Público (SIAP), que processam os arquivos de maneira diferenciada,

sem possuir entre elas qualquer. Este fato foi solucionado com a reengenharia de processos e processos de carga bem definidos, ficando evidente a importância dos processos ETL na resolução de desafios de transformação e integração de dados.

- **Armazenamento dos Dados – *Loading***

É o ultimo passo no processo ETL, e tem por função carregar os dados no *data warehouse*, para que assim, se tornem acessíveis aos usuários e ferramentas. De acordo com Chaudhuri (1997), o carregamento dos dados pode ser realizado de duas maneiras:

- ❏ *Refresh* – Os dados armazenados são completamente reescritos.
- ❏ *Update* – Somente as mudanças aplicadas a base de dados é adicionado ao *data warehouse*. Normalmente não são utilizadas operações de exclusão e modificação.

2.1.4.3 Área de Apresentação dos Dados

A área de apresentação dos dados é local onde os *data marts* e o *data warehouse* se encontram. Sendo assim, os dados estão armazenados e organizados de maneira que usuários e ferramentas analíticas possam acessá-los por meio de queries.

Esses dados estão estruturados em *schemas* multidimensionais, sendo o mais conhecido entre eles, o *star schema*. Tais conceitos serão tratados mais profundamente na seção 2.6, Modelagem Multidimensional.

2.1.4.4 Ferramentas de Acesso aos Dados

Por último, o ambiente que é utilizado pelo usuário final, orientado por ferramentas, que permitem a construção e execução de *queries* com a finalidade de obter informações úteis para a tomada de decisão. Por definição, todas as ferramentas de acesso processam os dados localizados na área de apresentação dos dados do *data warehouse* (Kimbal, 1996).

Diversas ferramentas podem ser utilizadas para acesso, desde as mais simples até as mais complexas, como em alguns casos de mineração de dados. Entre outros

exemplos, podem se citar ferramentas de predição, análise, modelagem e *ad hoc query*.

2.1.5 Modelagem Multidimensional

A principal característica de sistemas OLAP, componente de um *data warehouse*, é possuir multi dimensionalidade, ou seja, os dados pré-processados são estruturados em um formato de cubo, com o intuito de facilitar e agilizar as respostas requeridas pelas queries gerenciais. Sendo assim, a análise multidimensional permite que usuários acessem um grande número de fatores interdependentes e visualizem as complexas relações entre tais dados (Ballard *et al*., 1998).

Tal metáfora pode ser explicada pelo fato do cubo de dados permitir que a informação seja modelada e visualizada em múltiplas dimensões, ou seja, cada parte do cubo representa uma dimensão. Por exemplo, os dados de uma loja com várias filiais podem ser armazenados em um cubo de 3 dimensões, sendo elas:

- ❑ Cliente (ClienteID, ClienteNome)
- ❑ Loja (LojaID, LojaNome)
- ❑ Tempo (Ano, Mes, Hora, Min)

Baseado neste exemplo, a Figura 7 representa um cubo de dados de um d*ata warehouse*, possuindo as 3 dimensões citadas anteriormente (Cliente, Loja e Tempo). Além dessas dimensões, outros atributos também podem fazer parte do cubo.

Existem dois tipos de tabelas que compõe um modelo multidimensional, sendo elas as tabelas de fatos e as tabelas de dimensão, possuindo diferentes características. A seguir, definem-se cada uma individualmente.

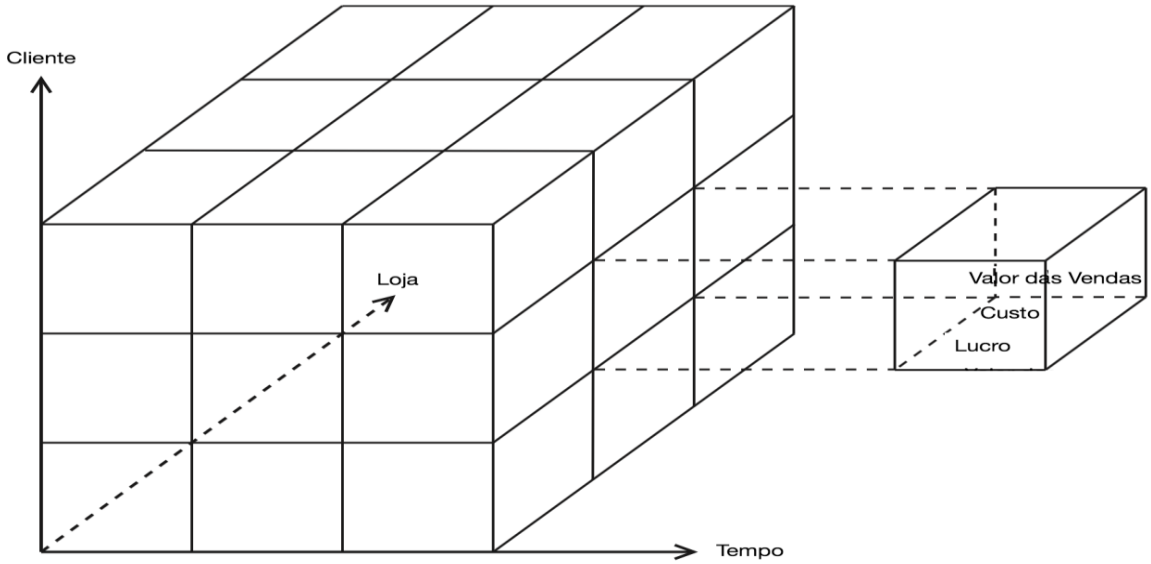


Figura 7: Cubo de dados de um *data warehouse*. Fonte: Rainardi (2008).

2.1.5.1 Tabela de Fatos

As tabelas de fatos devem conter valores numéricos de medição, ou seja, fatos representam medições de desempenho por meio de números, como por exemplo: soma de gastos, número total de clientes e número total de itens vendidos. Além disso, as tabelas de fatos possuem chaves estrangeiras (*Foreign Keys*), que associam os fatos com as tabelas de dimensão. Sendo assim, todas as tabelas de fato possuem duas ou mais chaves estrangeiras, já que a tabela de fatos em si possui sua própria chave primária, gerada de um *subset* das chaves estrangeiras, sendo chamada de chave composta ou concatenada. Portanto, toda tabela que possuir chave composta, é uma tabela de fato (Kimbal, 1996).

De acordo com Ballard *et al*. (1998), as tabelas de fato possuem as seguintes características:

- Contém um pequeno número de colunas.
- Possuem grandes quantidades de linhas em comparação com as tabelas de dimensão.

- As informações são normalmente numéricas.
- Os dados devem possuir características aditivas ou semi-aditivas.

Um exemplo de tabela de fatos está representado na Figura 8.

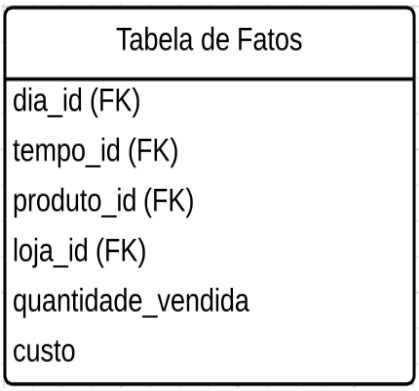


Figura 8: Exemplo de tabela de fatos. Fonte: Autoria Própria.

2.1.5.2

Tabela de Dimensões

Dimensões são descrições textuais que descrevem os fatos, ou seja, participam na definição de detalhes de um fato. Normalmente, são descritivas, por exemplo:

- Língua (Português, Inglês, Francês...).
- Gênero (Masculino, Feminino).
- País (Brasil, Alemanha, África...).
- Produto (Nome, Descrição…).

As tabelas de dimensões ajudam a entender os números contidos nas tabelas de fatos, auxiliando assim a interpretação dos dados. Segundo Ballard *et al*. (1998), essas tabelas normalmente possuem menos linhas do que as tabelas de fatos, por outro lado, possuem um número maior de colunas. A Figura 9 exemplifica a estrutura de uma tabela de dimensões.

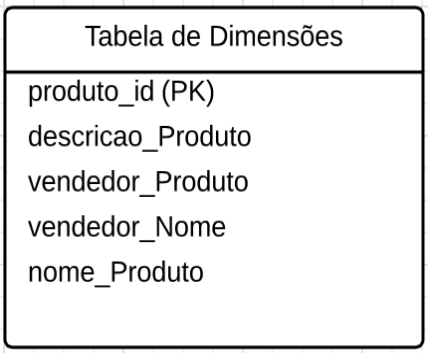


Figura 9: Exemplo de tabela de dimensões. Fonte: Autoria Própria

2.1.5.3 Tipos de Modelos Dimensionais

Na próxima seção são apresentadasduas formas mais conhecida de modelagem dimensional, sendo denominadas como modelo estrela e modelo floco de neve.

O modelo estrela tem como característica possuir uma tabela de fatos central, sendo cercada e conectada por meio de relacionamentos com diversas tabelas de dimensões. As tabelas de dimensões, menores, são as pontas da estrela, enquanto a tabela de fatos é o centro.

Este modelo se tornou muito comum pelo fato de prover alta performance em comparação com os modelos normalizados (E/R), que são associados com um banco de dados relacional. Além disso, pode-se citar que esse modelo é de fácil entendimento em relação aos demais (Ballard *et al.,* 2005). A Figura 10 representa um modelo estrela, possuindo uma tabela de fatos central e quatro tabelas de dimensão (*Date Dimension, Week Dimension, Product Dimension e Supplier Dimension*) com seus respectivos atributos.

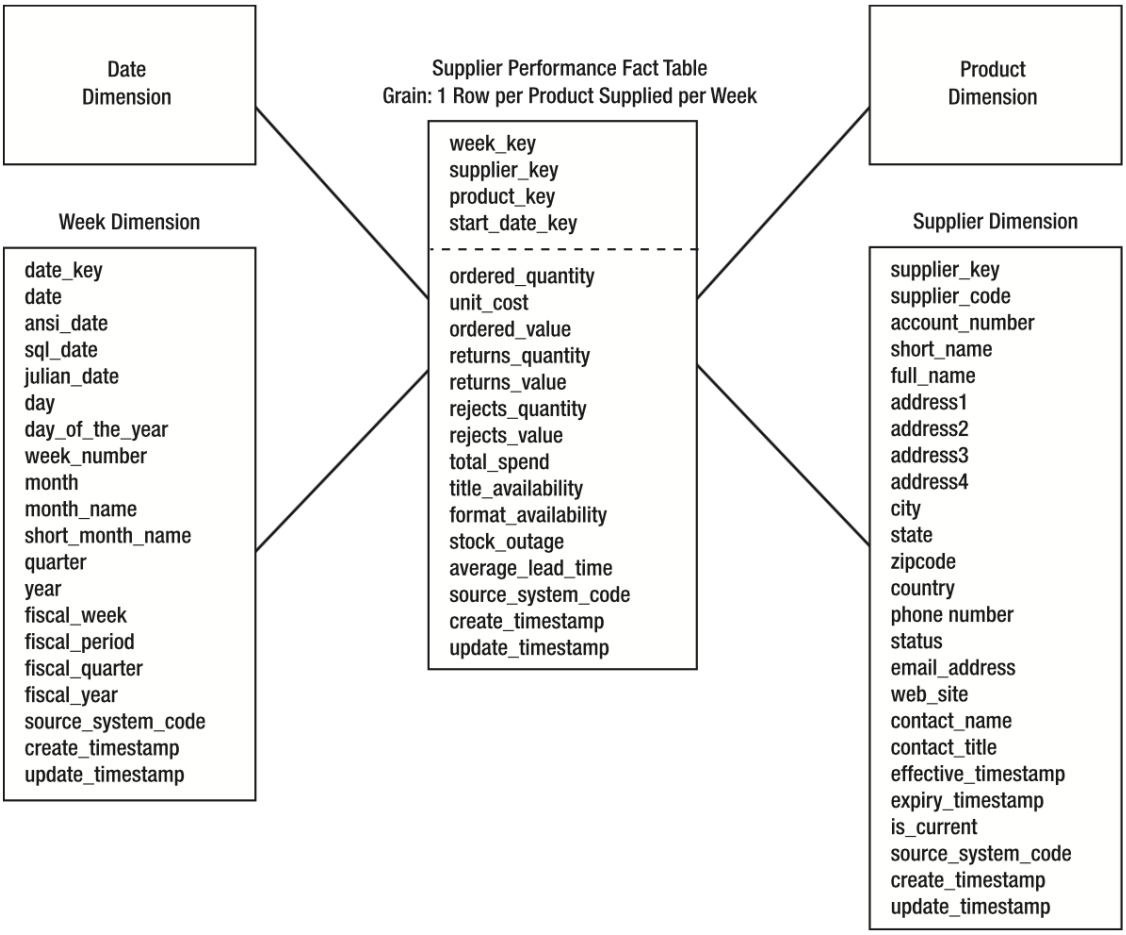


Figura 10: Modelo estrela. Fonte: Rainardi (2008).

Já o modelo floco de neve é similar ao modelo estrela, sendo que as tabelas de dimensão são normalizadas em tabelas relacionadas. O método utilizado para normalizar as tabelas de dimensão, é chamado de *snowflaking*, sendo que os atributos de baixa cardinalidade são retirados e separados em tabelas normalizadas (Ballard *et al.*, 2005). O grande problema do modelo floco de neve é a dificuldade de utilização devido a quantidade de *joins* e também por apresentar pior desempenho em comparação ao modelo estrela.

2.1.6 Índices

A capacidade de acesso rápido e descomplicado a dados demonstram a essência de um *data warehouse*, caso as informações não apresentem um nível eficiente e facilitado de indexação, o *data warehouse* não será um sucesso (Inmon, 1996). Neste contexto, Golfarelli *et al.* (2009) aborda que o processo de definição e seleção de índices, tornou-se um dos principais temas de busca em *data warehouse*, pois apresentam técnicas eficazes para melhorar o desempenho do sistema.

Por sua vez, Kimbal (1996) apresenta índices como a espinha dorsal do *data warehouse* quando o assunto é tempo de resposta em consultas, porém salienta que em contra partida aos benefícios proporcionados na consulta, observa-se uma sobre carga na gestão de índices durante o processo de ETL. Abordando técnicas de indexação, Imhoff *et al*. (2003) apresenta dois tipos básicos que podem ser utilizados ao modelar um *data warehouse*: árvore-B e bitmap, apresentados na sequência.

2.1.6.1 Árvore-B

Tratando-se de índices específicos, Johnson *et al.* (2008) apresenta a estrutura de dados árvore-B, conforme o conceito de um índice de um livro, utilizando analogia de árvore para armazenar dados em nós pai e filho, conforme Figura 11.

Neste assunto, Imhoff *et al*. (2003) complementa que índices árvore-B utilizam uma estrutura de árvore recursiva para armazenar o valor do índice e ponteiros para outros nós, assim a cada nó visitado exige-se uma decisão binária, comparando, por exemplo, se o valor do índice é maior ou menor que o valor buscado, conforme estrutura ilustrada na Figura 12.

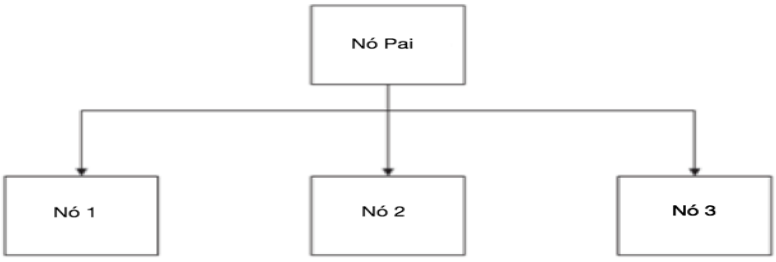


Figura 11: Modelo conceitual de índices árvore-B. Fonte: Johnson *et al.*, (2008).

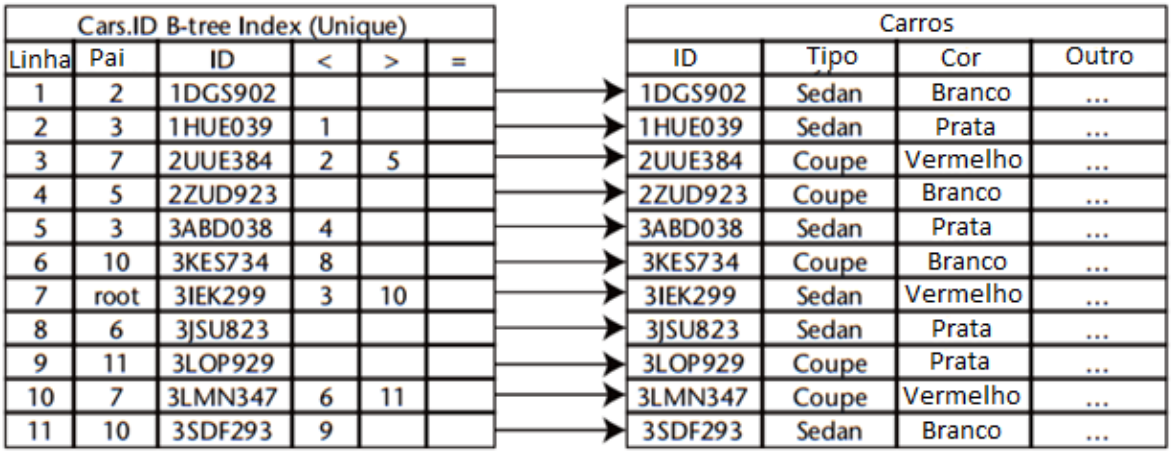


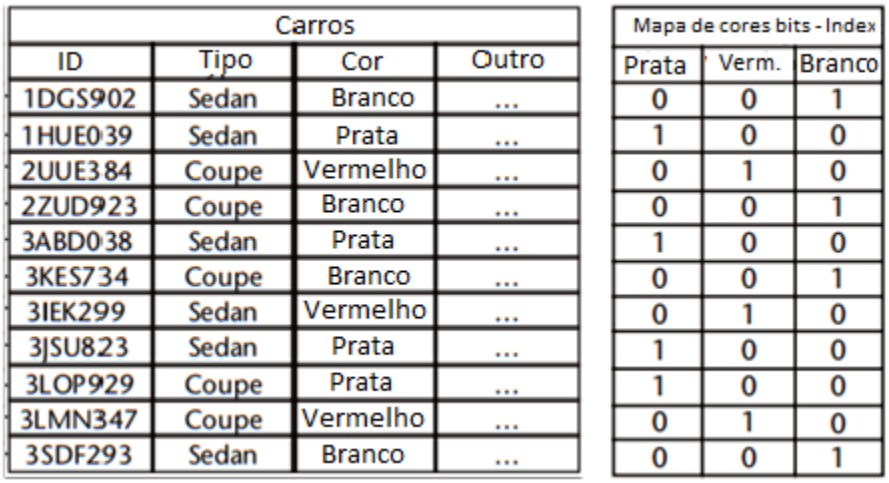
Figura 12: Estrutura de índices árvore-B. Fonte: Imhoff *et al.,* (2003).

Contextualizando a utilização do índice arvore-B, (Kimbal, 1996) sugere a utilização deste tipo de índice em colunas de atributos de alta cardinalidade utilizada para restrições. Deste modo Imhoff *et al*. (2003) complementa que índices do tipo árvore-B demonstram-se a melhor escolha em consultas onde os caminhos são conhecidos e controlados, e são extensivamente utilizados na modelagem de bases OLTP.

2.1.6.2 Mapa de Bits

Conceituando mapa de bits, Inmon (1996) explica que se trata de uma forma especializada de um índice capaz de indicar a existência ou não, de uma condição para um grupo de blocos ou registros.

Neste contexto, Ciferri (2002) apresenta que o índice do tipo mapa de bits, tratando-se de um atributo indexado, consiste em um vetor de bits que armazena valores binários, com finalidade booleana ao predicado considerado. Deste modo, Imhoff *et al*. (2003) exemplifica na Figura 13, a criação de um índice do tipo mapa de bits, para cores de veículos.



Carros			
ID	Tipo	Cor	Outro
1DGS902	Sedan	Branco	...
1HUE039	Sedan	Prata	...
2UUE384	Coupe	Vermelho	...
2ZUD923	Coupe	Branco	...
3ABD038	Sedan	Prata	...
3KES734	Coupe	Branco	...
3IEK299	Sedan	Vermelho	...
3JSU823	Sedan	Prata	...
3LOP929	Coupe	Prata	...
3LMN347	Coupe	Vermelho	...
3SDF293	Sedan	Branco	...

Mapa de cores bits - Index		
Prata	Verm.	Branco
0	0	1
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1
0	1	0
1	0	0
1	0	0
0	1	0
0	0	1

Figura 13: Índice mapa de bits. Fonte: Imhoff *et al.,* (2003).

Relacionando a utilização de mapa de bits, Kimbal (1996) sugere que para colunas com número limitado de valores (baixa cardinalidade), torna-se a técnica mais apropriada. Já Inmon (1996) destaca que mapas de bits possuem alto custo de construção e manutenção, mas fornecem facilidade de acesso e comparação veloz de registros.

2.1.7 Etapas do Projeto

Tratando-se de um projeto de *data warehouse*, Kimbal (1996) enfatiza que o foco do projeto deve ser a necessidade do negócio, e destaca que durante o processo de criação do ambiente, cada etapa deve possuir um ciclo finito, com início e fim definidos. Seguindo neste conceito, Rainardi (2008) apresenta uma estrutura em cascata das etapas de um projeto, conforme Figura 14, e complementa que para produção de um projeto qualificado, é necessário além de experiência, conhecer profundamente cada etapa, sendo possível estimar custos e prazos.

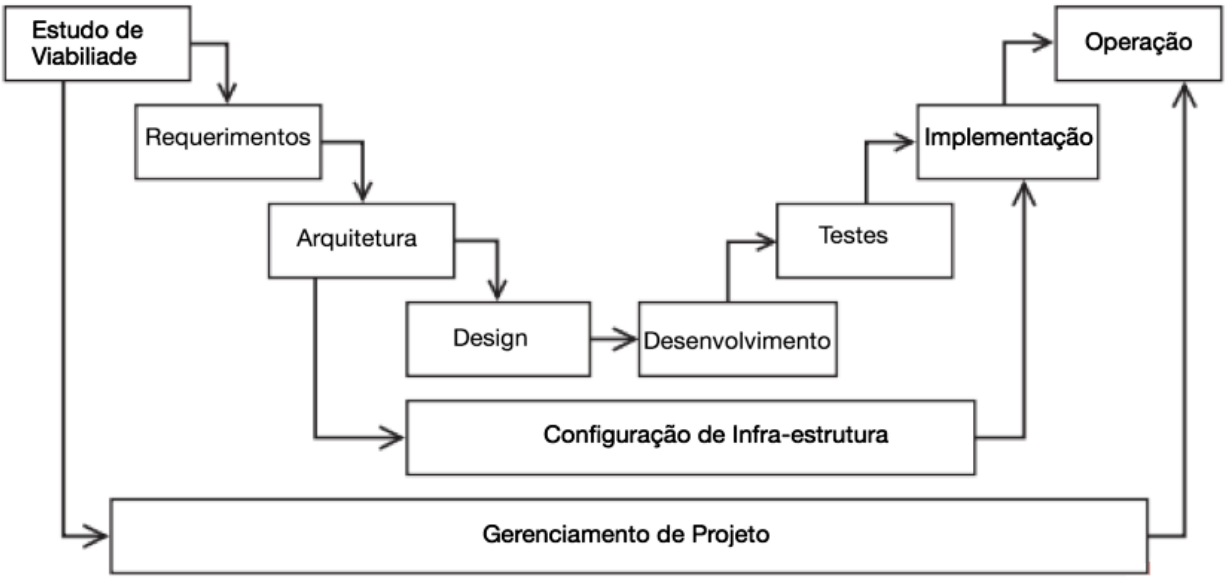


Figura 14: Etapas de um projeto. Fonte: Rainardi (2008).

Deste modo, Rainardi (2008) apresenta as etapas abordadas na figura 14, da seguinte forma:

- Estudo de Viabilidade: nesta etapa verificam-se requisitos em alto nível, como a necessidade sobre um *data warehouse* e por qual motivo este seria a solução, realiza-se uma breve verificação sobre os sistemas envolvidos e uma análise de prazos e custos.
- Requerimentos: nesta etapa realiza-se uma conversa com os usuários para compreender os detalhes dos processos, o negócio, os dados e as questões.
- Arquitetura: esta etapa consiste basicamente em determinar qual arquitetura de fluxo de dados e sistema que será utilizada, incluindo a especificação para servidores de banco de dados, o tipo de rede, a solução de armazenamento, e assim por diante.
- Modelagem: nesta etapa realiza-se a modelagem de três partes principais do *data warehouse*: os armazenamentos de dados, o sistema de ETL, e as aplicações *front-end*.
- Desenvolvimento: esta etapa consiste na criação das três partes modeladas.

- Testes: nesta etapa realiza-se o teste do armazenamento de dados, do sistema de ETL, e as aplicações *front-end,* sendo a etapa responsável por buscar e solucionar problemas no ambiente criado.
- Implantação: após a conclusão do sistema, este é o momento para colocar o *data warehouse* em produção, realizar a primeira carga, orientar usuários e o time de operação, criar guias para utilização e suporte do sistema.
- Operação: neste momento, o time de operação continua com a administração do *data warehouse,* resolvendo erros e problemas e administrando novos usuários e o controle de acesso.
- Configuração de infraestrutura: esta é uma das maiores etapas do projeto, consiste na criação da arquitetura do sistema, modelagem técnica, compra de hardware e software, instalação do hardware e software, configurações de rede, teste de infraestrutura e produção de documentações.
- Gerenciamento do projeto: etapa responsável pelo planejamento e controle das atividades envolvidas no projeto, também é responsável pela comunicação entre as equipes para resolução de problemas.

Neste contexto, Ciferri (2002) enfatiza a importância da definição da arquitetura para um projeto, e destaca as seguintes atividades:

- Seleção de servidores para armazenamento, banco de dados, OLAP e também ferramentas clientes;
- Integração de servidores e ferramentas clientes;
- Identificação dos provedores de informação, dados a serem armazenados e sua integração ao ambiente;
- Definição da organização física do *data warehouse* e escolha de métodos de acesso.

2.1.8 Softwares: Vantagens e Aplicações

Corporações que utilizam aplicações de *data warehousing* possuem uma natureza distribuída, deste modo vantagens da utilização de uma arquitetura distribuída tornam-se muito mais expressivas, como o suporte a um número elevado de usuários,

aumento da capacidade de processamento de consultas, confiabilidade e disponibilidade do sistema (Ciferri, 2002).

Abordando a escolha do *software* adequado, Rainardi (2008) enfatiza que a escolha de software na construção do *data warehouse*, afeta diretamente a arquitetura do sistema, pois versões específicas do SQL Server, Oracle, ou Teradata por exemplo, possuem requisitos de arquitetura diferentes. Johnson *et al.* (2008) intensifica que a escolha do software adequado, depende de especificações do ambiente e requisitos.

Mundy *et al.* (2008), sugere como uma vantagem do SQL Server, a opção de ferramentas disponíveis para manter tabelas agregadas no banco de dados relacional, visto que pessoas utilizam *views* indexadas, como um substituto para tabelas agregadas.

Imhoff *et al*. (2003), aborda algumas vantagens dos SGDBs Oracle, entre elas, a capacidade de trabalhar com estrutura recursiva em instruções de seleção de dados, possibilitando que ferramentas OLAP trabalhem com hierarquias armazenadas como estruturas de árvore recursiva e também a possibilidade de criar ambientes multilinguais, por meio da utilização de sinônimos em esquemas de dados diferentes.

Contextualizando o ambiente de software livre, Almeida (2004) enfatiza que dentre os SGBDs de código aberto o PostgreSQL é considerado o mais robusto e demonstra maior maturidade em comparação com outros SGBDs de código aberto, e destaca ainda que a comunidade de código aberto busca tornar o PostgreSQL o mais abrangente possível para as mais diversas plataformas e sistemas operacionais.

2.1.9 Trabalhos Relacionados

Os problemas e soluções de construção de um *data warehouse* são relatados em diversos outros estudos. A seleção dos trabalhos relacionados que serão tratados nesta seção foi motivada por: i) definirem conceitos importantes na área de data warehouse, ii) tratarem sobre a qualidade e integridade dos dados, iii) utilizarem metodologias ou processos que também serão implementados neste projeto e finalmente iv) apresentarem propostas de DW para órgãos públicos.

Santos *et el.* (2006), apresenta um estudo de caso na criação de um DW para a Secretaria de Saúde Pública de São Paulo. O projeto foi estruturado

organizacionalmente com níveis de hierarquia, começando no comitê executivo até a equipe de TI. A técnica abordada foi a *Source-Driven*, onde os requisitos são identificados pelos sistemas provedores dos dados que serão integrados e inseridos no *Data Warehouse*. Também foi desenvolvido um modelo relacional, com a finalidade de facilitar o processo de carga e permitir consultas operacionais. Já para as ferramentas escolhidas, alguns aspectos foram analisados, como: robustez para suportar o volume de dados, custo acessível e casos de sucesso. Dentre elas, podemos citar: Oracle 10g; Oracle IAS; OWB; Oracle Warehouse Builder e Compucarga. Por fim, a metodologia utilizada foi a *star schema*. É importante citar, que as maiores dificuldade enfrentadas pelos autores foi a falta de comunicação entre as equipes de negócio e de TI, além da compreensão sobre os processos mais detalhadamente.

Hu (2010) discute em seu trabalho o desafio de facilitar e tornar viável a troca de dados entre os mais variados sistemas e departamentos do governo na cidade de Nanhai, China. Para resolver tal problema, foi necessário estabelecer um DW baseado em uma arquitetura composta por seis elementos: 1) plataforma de compartilhamento e troca de dados 2) *kernel database*, 3) plataforma de suporte à aplicação, 4) aplicação do banco de dados, 5) plataforma de gerenciamento do banco de dados central e 6) plataforma de segurança. Os sistemas alimentadores deste data warehouse seriam departamentos de polícia, educação entre outros. XML foi o formato adotado como padrão para estruturação os dados, sendo que um dos motivos é a facilidade de transformar dados de certa plataforma em XML e integrar este arquivo convertido com outros sistemas. O método proposto pelo autor foi implementado e já está em uso em Nanhai. Futuramente, esta solução também pode ser útil e adaptável ao cenário de Curitiba.

Mussi *et al.* (2004), apresenta como transcorreu a implantação de um *data warehouse* na Agência Nacional de Vigilância Sanitária (ANVISA), motivada pela demanda de informações não estruturadas para a tomada de decisão. As principais fontes de dados do *data warehouse* são os sistemas transacionais (OLTP) que atendem aos usuários da ANVISA e do Ministério da Saúde. Esses sistemas possuem

bases de dados relacionadas com o assunto de interesse e que estão disponíveis para os processos de extração, transformação e carga (ETL). Os dados carregados no DW estão ao nível de granularidade mais atômica possível e de forma corporativa incremental e a modelagem dos dados é multidimensional, na visão estrela (*star schema*). No decorrer do projeto, Mussi *et al.* (2004) relaciona alguns problemas como: falta de delimitação do escopo e previsão orçamentaria compatível, baixo comprometimento dos profissionais envolvidos, o processo de ETL serviu para corrigir imperfeições dos provedores de informação e problemas relacionados com as características dos dados. Apesar destes problemas, estudos realizados pós implantação, mostram que é possível, a partir do DW implantado, obter diversos relatórios gerenciais, com a especificação de cada usuário, para que este possa tomar a melhor decisão.

Clemes (2001) apresenta a proposta de criação e implementação de um modelo de *data warehouse* que contemple o fornecimento de informações precisas e consistentes voltados para processos decisórios no âmbito de uma Instituição de Ensino Superior (IES). A arquitetura e o roteiro são testados em uma aplicação desenvolvida na UFSC. Neste trabalho observam-se contribuições significativas para futuras iniciativas em construção de *data warehouse*, o autor destaca problemas relacionados com os provedores de informação no ambiente universitário, muitos dados desejados pelos usuários não estão disponíveis nos sistemas operacionais da organização e ainda muitos departamentos utilizam soluções próprias para atividades comuns, potencializando este problema.

Deste modo destaca-se também, a importância da escolha da metodologia de desenvolvimento e a definição da infraestrutura de suporte ao ambiente. A eficácia da solução apresentada foi comprovada por meio do desenvolvimento do protótipo do sistema de suporte a decisão na UFSC, reunindo num só ambiente, informações de diversas fontes.

2.2 Governo Eletrônico

De modo geral, governos são responsáveis por uma quantidade significativa de informações para uso em suas operações internas e prestação de serviços (Diniz,

2009). Atualmente, esses dados estão cada vez mais disponíveis para a população, com a finalidade de contribuir para a transparência (Araújo, 2011). Essa prática deu origem ao chamado Governo Eletrônico, conhecido mundialmente como *E-Government*, que permite o cidadão acessar dados públicos mais rapidamente, de maneira democrática e eficiente (Hu, 2010).

Dados públicos governamentais tem o potencial de melhorar a tomada de decisão, assim como facilitar para as partes interessadas o acesso total e livre aos dados públicos além de abrir a oportunidade para as pessoas a avaliar o desempenho de várias instituições administrativas (UNPAP, 2014). Ou seja, introduz uma nova abordagem para a publicação de dados de governos e ajuda a preencher a lacuna entre governo e cidadãos.

De acordo com o relatório *United Nations E-Government* 2014, produzido pelo *United Nations Public Administration Programme* (UNPAP), juntos, o *E-Government* e a inovação podem proporcionar oportunidades significativas para transformar a administração pública em um instrumento de desenvolvimento sustentável, além de permitir que agências do governo centralizem a tomada de decisão.

Este relatório é produzido a cada dois anos, e uma de suas finalidade é avaliar a situação de desenvolvimento de 193 países desde 2003, verificando as seguintes dimensões: (i) a disponibilidade de serviços on-line, (ii) infraestrutura de telecomunicações e (iii) capacidade humana. Mais especificadamente, descreve a situação atual do *E-Government* em cada país e evidencia os benefícios de governo eletrônico para o desenvolvimento sustentável.

Essa iniciativa reforça a importância, além de destacar pontos críticos afetados pela integração e colaboração entre agências do governo de cada país, dos quais podem se citar algumas recomendações:

1. Não existem um único ministério ou departamento do governo que pode efetivamente lidar com questões e unir todas as informações referentes à erradicação da pobreza, por exemplo, sendo um problema causado por diversas causas e variáveis. Desta maneira, a colaboração entre órgãos públicos é necessária para melhor entendimento e tratamento de tal fato.

2. Com o avanço das tecnologias, a população demanda busca por soluções de maneira rápida e eficiente, assim como a prestação de contas, relatórios entre outros. Sendo assim, é exigido por parte do governo uma mudança na estrutura, apto à responder mudanças constantes, o que pode resultar da integração de serviços.
3. Devido à demanda dos cidadãos por uma participação mais significativa nos assuntos públicos e tomada de decisão, é exigido por parte do governo mecanismos que permitam à população meios de envolvimento nas decisões que afetam suas vidas. Além do mais, pessoas não ligadas ao governo podem estar envolvidos na criação de serviços e soluções para os desafios sociais.
4. Instituições, sistemas governamentais e processos precisam se adaptar rapidamente, através de uma gestão eficaz do conhecimento a nível de governo tanto local quanto nacional.

A Figura 15, mostra a classificação do Brasil (57º) de acordo com a capacidade e a utilização das práticas de *E-Government.*

País	Nível de Renda	EGDI	2014 Rank	2012 Rank	Mudança no Rank
		EGDI Muito Alto			
United States of America	Alto	0.8748	7	5	↓ 2
Canada	Alto	0.8418	11	11	-
		EGDI Alto			
Uruguay	Alto	0.7420	26	50	↑ 24
Chile	Alto	0.7122	33	39	↑ 6
Argentina	Acima do Intermediário	0.6306	46	56	↑ 10
Colombia	Acima do Intermediário	0.6173	50	43	↓ 7
Costa Rica	Acima do Intermediário	0.6061	54	77	↑ 23
Brazil	Acima do Intermediário	0.6008	57	59	↑ 2
Barbados	Alto	0.5933	59	44	↓ 15
Antigua and Barbuda	Alto	0.5927	60	49	↓ 11
Mexico	Acima do Intermediário	0.5733	63	55	↓ 8
Venezuela (Bolivarian Republic of)	Acima do Intermediário	0.5564	67	71	↑ 4

Figura 15: Ranking mundial – *E-Government*. Fonte: UNPAP (2014).

De acordo com Araújo (2001) existem diversos fatores que são responsáveis pela posição atual do Brasil, tais como a insuficiência de serviços online e a deficiente infraestrutura de telecomunicações.

Já a China, teve sua primeira iniciativa de E-Government no final de 1980, onde os governos ambos de nível central e local construíram um sistema de automação de escritórios, além de estabeleceram uma intranet. No começo de 1990, a China já possuía cinco projetos com foco em E-Governement.

De acordo com Hu (2010), com a finalidade de acelerar o ritmo da mudança em funções do governo para atender a exigência de reforma, abertura e modernização das políticas, melhorar o desempenho de operação do governo, introduzir novas medidas governamentais de forma científica e mecanismos mais eficazes para acompanhar as atividades econômicas e prestar um melhor serviço para o público, a China decidiu implementar quatro base de dados, utilizando a tecnologia *data warehouse*. São elas:

1. Informações básicas da população;
2. Informações básicas jurídicas;
3. Informações básicas de geografia e recursos naturais e
4. Informações básicas sobre a economia.

2.2.1 Dados Abertos no Brasil

No Brasil, dentre os órgãos públicos que utilizam a tecnologia do *data warehouse*, os mais relevantes para esta pesquisa, que serão analisados mais profundamente, são os seguintes: Serpro[3], Tesouro Nacional (Santos, 2011), Secretaria da Receita Federal[5], Tribunal de Contas da União[5].

Conhecido como Serviço Federal de Processamento de Dados, o Serpro é uma empresa vinculada ao ministério da fazenda, responsável por prestar serviços em tecnologia da informação, a fim de evidenciar e promover a transparência e controle dos gastos públicos. Além disso, o Serpro é responsável por administrar um *data warehouse* para o Governo Federal[5], possuindo total domínio sobre as informações

[5] https://www.serpro.gov.br/sobre/a-empresa - Acesso em 12.05.2015

contidas neste armazém de dados, com a finalidade de integrar diversas bases e facilitar a tomada de decisão e estratégico do Governo Federal[6].

Em Março de 2006 o TCU também firmou um contrato com o Serpro, para produzir e implementar o sistema Síntese[7] (Sistema de Inteligência e Suporte ao Controle Externo), com um custo total de 12,5 milhões de reais[8]. Assim como os demais, o objetivo desse sistema era a criação de um *data warehouse* e a unificação das informações obtidas de fontes distintas. Algumas das vantagens com a implementação deste *data warehouse*, seriam:

- A melhoria do planejamento das ações de controle;
- A possibilidade de detecção de indícios de fraude por meio de tratamento estruturado de dados e
- o controle externo eletrônico.

Já na Secretaria da Receita Federal (SRF)[9], primeiramente foi realizado um processo de levantamento de dados e requisitos necessários para a o planejamento do modelo do negócio, com o intuito de implementar o *data warehouse*. Este modelo é composto por dimensões como: Tributo, Atividade econômica, natureza Jurídica, Organização SRF, Localização, Tempo, Situação do Contribuinte. É importante citar que para popular o armazém de dados da SRF, foi necessário realizar etapas de identificação, avaliação e definição de modelos, levantamento de requisitos de qualidade e identificação das fontes de dados[10].

Além dos *data warehouse* citados anteriormente, o Serpro ainda é responsável por uma vasta gama de servidores de órgãos públicos, que são mostrados na Figura 16, com seus respectivos nomes, clientes, volume em *gigabytes* e a quantidade de usuários.

[6] http://www4.serpro.gov.br/imprensa/publicacoes/tema-1/tematec/1996/ttec27/ - Acesso em 12.05.2015
[7] http://www4.serpro.gov.br/imprensa/publicacoes/tema-1/antigas%20temas/tema_175/materias/informacoes-lapidadas/ - Acesso em 02.12.2015
[8] http://www4.serpro.gov.br/noticias-antigas/noticias-2006/20060320_07 - Acesso em 31.05.15

[9] http://idg.receita.fazenda.gov.br - Acesso em 02.12.2015
[10] http://www1.serpro.gov.br/publicacoes/tema/162/materia14.htm - Acesso em 02.12.2015

Sendo assim, na página do Serpro é possível encontrar dados e informações, provindas de diversos órgãos públicos, com a finalidade de facilitar o acesso à informação, conforme determina a Lei de Acesso à Informação (Lei 12.527, de 18/11/2011) que garante a sociedade o acesso às informações produzidas ou custodiadas pelo poder público e não classificadas como sigilosas.

Dentre as informações disponibilizadas, encontram-se:

- Informações referentes à prestação de contas do Serpro, por exercício fiscal.
- Dados sobre execução orçamentaria e financeira do Serpro, contemplando diárias e passagens, despesas e grandes obras.
- Informações sobre licitações, consultas públicas, contratos e contratações firmados pelo Serpro.
- Informações sobre os empregados públicos em exercício no Serpro e concursos públicos de provimento de cargos.
- Dados sobre programas e ações da empresa que possuem relações com a sociedade.
- Estrutura organizacional, competências, telefones e endereços, informações sobre funções e agenda de autoridades, relação de clientes.

Em Produção:			
Projeto	**Cliente**	**Volume (GB) (Grandeza)**	**Usuários (Aprox.)**
DW Corporativo SRF	RFB	4.000	3.200
ICOMEX	MDIC	300	180
RENACH	DENATRAN	120	10
RENAINF			30
SIG/PGFN	PGFN	100	300
SIAPE DW	MPOG	**670**	**2.500**
SIASG DW		100	200
SSDSPU		20	300
SINDEC	DNIT	3	200
SINTESE	TCU	*2.000*	*100*
BGU	STN	*200*	*10*
Pagamento Efetivo	STN	150	20
Em Desenvolvimento			
DW SNCR	INCRA	300	
DW Fiscal MTE	MTER	200	
Gestão Financeira	SERPRO	100	50
Aquaviário	MT		
Trânsito	DENATRAN		

Figura 16: Projetos de *data warehouse* - Serpro [11]

2.2.2 Dados Abertos em Curitiba

Centralizando o contexto para o município de Curitiba, observa-se um grande número de iniciativas por parte de órgãos públicos em disponibilizar dados para livre utilização da sociedade, apresentam-se na sequência as principais iniciativas neste contexto.

2.2.2.1 Instituto Paranaense de Desenvolvimento Econômico e Social

O IPARDES é responsável por gerenciar a Base de Dados do Estado (BDE*web*)[12], sendo este um sistema de informações estatísticas com mais de 9 milhões de dados classificados por grandes temas e assuntos. Nesta base, é possível obter informações das áreas física, econômica, social, financeira, política e administrativa do Estado do Paraná.

A consulta de informações na BDEweb, é segmentada de acordo com as seguintes características:

[11] http://www3.tesouro.fazenda.gov.br/Sistema_Informacao_custos/downloads/5_SERPRO_Seminario_Sistema_Custos.pdf - Acesso em 02.12.2015
[12] http://www.ipardes.pr.gov.br/imp/index.php - Acesso em: 27 maio. 2015.

- Localidade: apresenta o Estado do Paraná e suas divisões regionais e municipais;
- Variável: assunto escolhido para obtenção de dados;
- Período: utiliza-se periodicidade anual, permite-se a escolha de intervalos.

Este método de consulta apresenta o resultado exemplificado pela Figura 17, que pode ser visualizado via tabela em HTML ou obtido por meio de um arquivo CSV.

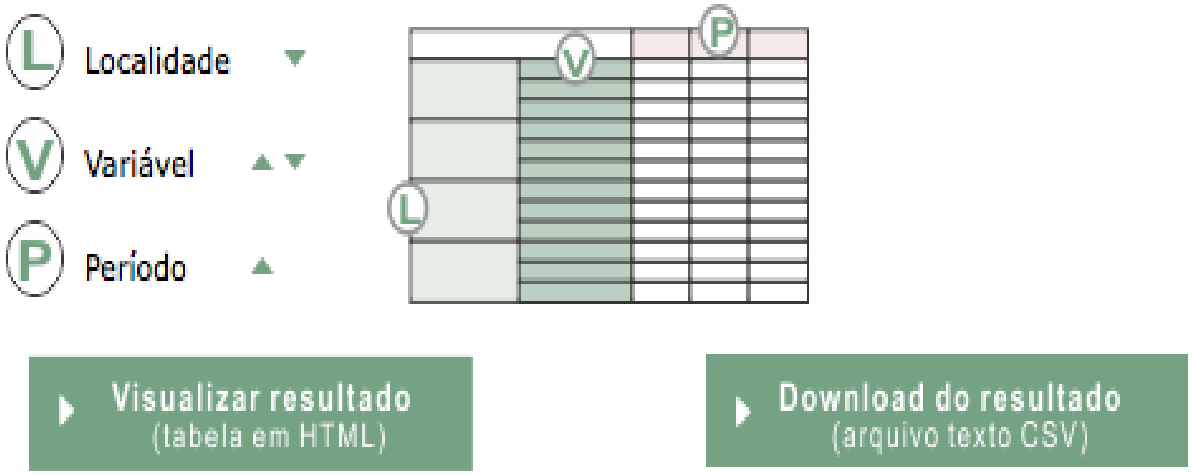


Figura 17: Estrutura do resultado obtido em consultas no BDEweb[13]

Seguindo neste modelo, apresenta-se na Tabela 2, um trecho de uma consulta organizada por assunto, neste caso abastecimento de água, onde observa-se a utilização da estrutura apresentada na Figura 17, realizada na base disponibilizada pelo IPARDES, sendo esta uma fonte de dados a ser integrada neste trabalho.

[13] http://www.ipardes.pr.gov.br/imp/index.php - Acesso em: 27 maio. 2015.

Localidade	Variável	2008	2009	2010	2011	2012	2013	2014
Abatiá	Abastecimento de Água - Unidades Atendidas	...	2.049	2.171	2.163	2.191	2.242	...
Abatiá	Abastecimento de Água - Unidades Atendidas Residenciais	...	2.049	2.085	2.111	2.144	2.191	...
Abatiá	Abastecimento de Água - Ligações	...	2.049	2.171	2.163	2.191	2.242	...
Abatiá	Abastecimento de Água - Ligações Residenciais	...	...	...	...	...	...	...
Adrianópolis	Abastecimento de Água - Unidades Atendidas	1.046	1.066	1.101	1.131	1.164	1.226	1.264
Adrianópolis	Abastecimento de Água - Unidades Atendidas Residenciais	948	960	995	1.028	1.055	1.112	1.137
Adrianópolis	Abastecimento de Água - Ligações	1.034	1.048	1.079	1.108	1.137	1.194	1.228
Adrianópolis	Abastecimento de Água - Ligações Residenciais	938	944	975	1.007	1.030	1.082	1.104

Tabela 2: Consulta organizada por assunto: Abastecimento de água (IPARDES)[14]

Em uma análise preliminar dos dados disponibilizados, verifica-se a padronização da estrutura de dados apresentada na Figura 17, como uma ponte de destaque para esta base de dados, visto que esta será integrada ao *data warehouse* neste trabalho.

[14] http://www.ipardes.pr.gov.br/imp/index.php - Acesso em: 27 maio. 2015.

2.2.2.2 Instituto de Planejamento Urbano de Curitiba

Analisando o contexto de dados no munícipio de Curitiba, o IPPUC disponibiliza o Portal Curitiba em Dados[15]. Neste portal, verifica-se informações de diversos domínios diferenciados abrangentes ao munícipio, como: saúde, esporte, infraestrutura urbana, educação, cultura, finanças, segurança, geografia, entre outros.

Uma análise preliminar deste portal, mostra que os dados são disponibilizados nos formatos XLS e PDF, com estrutura facilitada para visualização, conforme Figura 18, que apresenta uma relação de domicílios particulares permanentes urbanos com abastecimento adequado de água, agrupados por áreas de expansão demográfica em Curitiba no ano de 2010.

Curitiba - Áreas de Expansão Demográfica	AED	Domicílios Particulares Permanentes Urbanos		
		Total	Com Abastecimento Adequado de Água	
			Absoluto	Relativo(%)
Mercês	4106902999005	4,681	4,681	100.00
Vila Izabel	4106902999010	3,777	3,777	100.00
Barreirinha	4106902999016	5,026	5,026	100.00
Alto da Rua XV/Jardim Social/Hugo Lange	4106902999028	5,768	5,768	100.00
Água Verde ZR4	4106902999043	5,496	5,496	100.00
Centro	4106902999001	14,576	14,569	99.95
Uberaba 2	4106902999047	12,696	12,685	99.91
CIC PI - Nossa Senhora da Luz	4106902999055	8,918	8,908	99.89
Rebouças	4106902999002	5,764	5,758	99.89
Água Verde ZR3	4106902999042	4,101	4,096	99.88
Capão da Imbuia	4106902999007	6,200	6,190	99.85
Bacacheri	4106902999012	7,107	7,090	99.76

Figura 18: Domicílios com abastecimento adequado de água[13].

Seguindo no IPPUC, identifica-se a disponibilização de dados georreferenciados de equipamentos urbanos de Curitiba, no formato *shapefile,* facilitando o processo de integração em um banco de dados por exemplo. Verifica-se nestes dados, informações de diferentes domínios abrangentes aos equipamentos urbanos do munícipio, como por exemplo: bibliotecas, escolas, hospitais, mercados, entre outros. A Figura 19 apresenta um trecho de uma consulta de escolas públicas de educação básica.

[15] http://curitibaemdados.ippuc.org.br - Acesso em: 27 maio. 2015.

Equipamento : Escola		Tipo Equipamento : Educação Básica		Dependência Administrativa : Particular	
	Nome	Telefone	Endereço	Bairro	Regional
1	A PROJEÇÃO	3296-2765	R. GOV. JORGE LACERDA,	GUABIROTUBA	Cajuru
2	ACESSO	3016-2682	R. MONS. CELSO,	CENTRO	Matriz
3	ACESSO	3016-2629	AV. MAL. FLORIANO PEIXOTO,	BOQUEIRÃO	Boqueirão
4	ADVENTISTA ALTO BOQUEIRÃO	3378-6274	R. BOM PASTOR,	ALTO BOQUEIRÃO	Boqueirão
5	ADVENTISTA BOA VISTA	3051--8620	R. FERNANDO DE NORONHA,	SANTA CÂNDIDA	Boa Vista
6	ADVENTISTA BOM RETIRO	30143734	R. CLETO DA SILVA,	BOM RETIRO	
7	ADVENTISTA BOQUEIRÃO	3014-3734	R. TEN. FRANCISCO FERREIRA DE SOUZA,	BOQUEIRÃO	Boqueirão
8	ADVENTISTA CENTENÁRIO	3226-3636	R. ARGÉLIA,	CAJURU	Cajuru
9	ADVENTISTA PORT ÃO	3051-8680	R. FR. GASPAR DA MADRE DE DEUS,	PORTÃO	Portão
10	ADVENTISTA SANTA EFIGÊNIA	3053---8636	R. PROF. GUILHERME BUTLER,	BARREIRINHA	Boa Vista
11	ADVENTISTA VILA SÃO PEDRO	3346-9676	R. ANTÔNIO RIBEIRO MACEDO,	XAXIM	Boqueirão
12	ADVENTISTA VISTA ALEGRE	3079-9969	R. VER. ANTENOR PAMPHILO DOS SANTOS,	VISTA ALEGRE	Santa Felicidade
13	AMPLA & AÇÃO	3327-0350	R. LAUDELINO FERREIRA LOPES,	PINHEIRINHO	Pinheirinho
14	ANCHIETA	3223-5433	TRAV. TOBIAS DE MACEDO,	CENTRO	Matriz
15	ANCHIETA - SUBSEDE I	3346-4548	R. PEDRO GUSSO,	CIDADE INDUSTRIAL	CIC
16	ANDRÉ LUIZ	3111-1703	R. TOBIAS DE MACEDO JÚNIOR,	SANTO INÁCIO	Santa Felicidade
17	ANJINHO MÁGICO	3365-1433	AV. PRES. AFFONSO CAMARGO,	CAPÃO DA IMBÚIA	Cajuru
18	ANJO DA GUARDA	3225-2633	R. TREZE DE MAIO,	SÃO FRANCISCO	Matriz
19	ATUAÇÃO	3274-6262	R. PROF. ULISSES VIEIRA,	SANTA QUITÉRIA	Portão
20	AYMORÉ	3276-6390	R. PE. DEHON,	BOQUEIRÃO	Boqueirão
21	BAMBINATA	3297-3933	R. DR. LEÃO MOCELLIN,	SANTA FELICIDADE	Santa Felicidade
22	BANDEIRANTES	3266-0881	R. CUIABÁ,	CAJURU	Cajuru
23	BASTOS MAIA	3286-7004	R. DIOGO MUGIATTI,	BOQUEIRÃO	Boqueirão

Figura 19: Escolas públicas de educação básica[16].

2.2.2.3 Prefeitura Municipal de Curitiba

Na Prefeitura Municipal de Curitiba (PMC), verifica-se uma iniciativa de transparência e cidadania, o Portal Dados Abertos Curitiba[17], com objetivo de disponibilizar por meio da internet, bases de dados de diversos órgãos do Governo Municipal de Curitiba, como: Abastecimento, Saúde, Recursos Humanos, Finanças, entre outros.

Neste portal, disponibilizam-se, documentos, informações e dados governamentais de domínio público com frequência de atualização mensal e espectro temporal em três meses. Os dados são disponibilizados no formato CSV e as especificações da tabela no banco de dados em uma planilha no formato XLSX, conforme figura 20.

Nome do Campo	Tipo	Tamanho	Descrição
P7	int	4	Chave da tabela do tipo auto incremento
Exercício_Idf	int	4	código do ano
Exercício	int	4	ano
Exercício_Descricao	varchar	MAX	descrição do ano
Data Empenho_idf	int	4	código do mês
Data Empenho	varchar	MAX	mês
Data Empenho_Descricao	varchar	MAX	descrição do mês
Data Liquidado_Idf	int	4	código do mês liquidado
Data Liquidado	varchar	MAX	mês liquidado
Data Liquidado_Descricao	varchar	MAX	descrição do mês liquidado
Data Pago_idf	int	4	código do mês pago
Data Pago	varchar	MAX	mês pago
Data Pago_Descricao	varchar	MAX	descrição do mês pago

Figura 20: Estrutura de tabelas em Dados Abertos Curitiba[15].

[16] http://ippuc.org.br - Acesso em: 17 junho. 2015.
[17] http://www.curitiba.pr.gov.br/dadosabertos - Acesso em: 27 maio. 2015.

Uma análise preliminar dos dados disponibilizados, mostra que estes utilizam uma estrutura padronizada, proporcionando maior facilidade no caso de integração destes dados em um *data warehouse*.

2.2.2.4 Desafios no Cenário de Curitiba

Enquanto a disponibilização de dados abertos e processos de integração têm diversas vantagens, como o objetivo em aprimorar a comunicação entre agências governamentais, facilitando assim acessar, entender e utilizar os dados públicos (DINIZ, 2009), os desafios para desenvolver, implementar e transformar os sistemas governamentais também são consideráveis.

De acordo com a UNPAP (2014), a colaboração entre agências do governo não é uma tarefa simples, e podem ser listados diversos problemas entre questões políticas, organizacionais e técnicas, como citado a seguir:

- Falta de confiança entre órgãos públicos;
- Estabelecimento de padrões;
- Utilização de aplicativos que sejam de fácil manutenção ao longo do tempo;
- Privacidade e segurança dos dados;
- Visões diferenciadas;
- Motivações diferentes ou até mesmo competitividade entre ministérios e agencias e
- Prioridades e valores diferentes.

Observa-se ao analisar dados disponibilizados pelas entidades abordadas, que estes utilizam formatos e padrões diferentes, determinados por cada entidade, de acordo com interesses e requisitos próprios. Deste modo, intensifica-se o desafio de integração de dados, para a criação de um *data warehouse* dito como modelo para órgãos públicos, padrões estes discutidos com maior profundidade na próxima seção.

2.2.3 Padronização de Dados

A falta de padronização de dados, é algo recorrente no âmbito da Tecnologia da Informação, o mesmo dado é muitas vezes produzido, gerenciado, utilizado e armazenado por diversos produtores de forma isolada, em formatos e padrões

próprios, com objetivo de atender única e exclusivamente às necessidades individuais de usuários específicos (Dornelles *et al.,* 2013).

Contextualizando este problema, as entidades governamentais, que produzem grande volume de informação essencial, a cada dia sofrem maior exigência para que publiquem dados de forma aberta, transparente e processável por máquina. Tais necessidades levaram a criação de inciativas voltadas ao âmbito de dados abertos.

Deste modo uma organização internacional de adesão voluntária, foi criada em 2011 com oito países fundadores: Brasil, Indonésia, México, Noruega, Filipinas, África do Sul, Reino Unido e Estados Unidos, a Parceria de Governo Aberto, cujo principal objetivo é assegurar o compromisso dos governos em promover a transparência, lutar contra a corrupção e fortalecer novas tecnologias para reforçar a governança (OPEN GOVERNMENT PARTNERSHIP, 2012).

Neste contexto, Lopes *et al., (2014)* aborda os oito princípios para dados governamentais, baseados no *Open Government Working Group*[18]*:*

1. Completo: Todos os dados públicos estão disponíveis, eles não estão sujeitos à privacidade, segurança ou controle de acesso válido regulamentado por lei.
2. Primário: Os dados são publicados na forma mais bruta possível, com o máximo nível de detalhe, não abreviado ou alterado.
3. Atuais: Os dados são fornecidos o mais rapidamente possível, a fim de preservar o seu valor.
4. Acessibilidade: Os dados são disponibilizados o mais amplamente possível e para a maior possível variedade de propósitos.
5. Operabilidade por máquina: Os dados são razoavelmente estruturados para permitir o processamento automatizado.
6. Acesso não discriminatório: os dados devem estar disponíveis a todos, sem a necessidade de identificação ou de registo.
7. Formatos não proprietários: Os dados devem estar disponíveis em um formato do qual ninguém tem controle exclusivo sobre.

[18] http://opengovdata.org - Acesso em: 01 junho. 2015.

8. Licença livre: Os dados não estão sujeitas a leis de direitos autorais, marcas, patentes ou segredo industrial. Restrições de privacidade razoável, de segurança e de controlo são permitidos desde que sejam regulados pelos estatutos.

De acordo com a UNPAP (2014), a utilidade, qualidade e acessibilidade da informação dependem diretamente do formato utilizado para a publicação de dados, de modo que mais pessoas podem participar e beneficiar de análise de dados que, por sua vez, pode contribuir para uma melhor definição de políticas.

Neste contexto, a Figura 21 apresenta um gráfico sobre o formato de dados disponibilizados em número de países que disponibilizam dados abertos, observa-se que 86 países fornecem dados em dados estruturados operáveis por máquina (por exemplo, planilhas), 56 em formatos não-proprietários (por exemplo, CSV), 24 países fornecem *Application Programming Interfaces* (APIs) e apenas 11 países fornecem dados em padrões abertos.

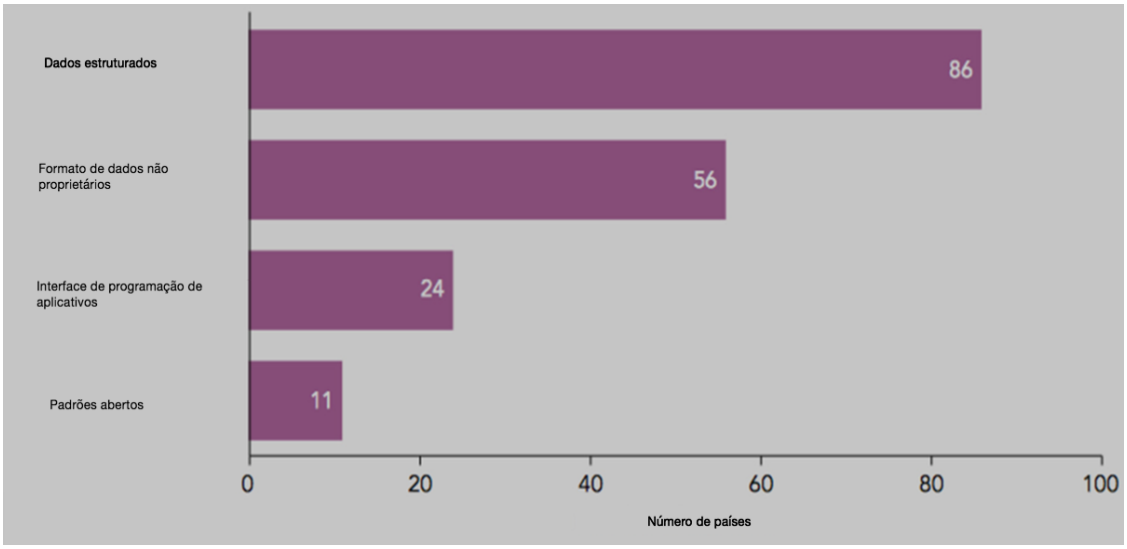


Figura 21: Número de países e formato de dados abertos. Fonte: UNPAP (2014).

2.4.2.1 Padrões de Interoperabilidade do Governo Eletrônico (e-PING)

Conforme Documento de Referência da e-PING (2015), os Padrões de Interoperabilidade do Governo Eletrônico (e-PING), objetiva-se em regulamentar a utilização da Tecnologia da Informação e Comunicação entre esferas do governo e sociedade em geral, por meio das seguintes premissas, políticas e especificações técnicas:

1. Adoção Preferencial de Padrões Abertos: Sempre que possível, serão adotados padrões abertos nas especificações técnicas. Padrões proprietários são aceitos nas seguintes condições:
2. Uso de Software Público e/ou Software Livre: A implementação dos padrões de interoperabilidade deve priorizar o uso de software público e/ou software livre.
3. Transparência: Com mais informações disponíveis é possível minimizar o número de interações do cidadão com o governo.
4. Segurança: A interoperabilidade na prestação dos serviços de governo eletrônico deve considerar o nível de segurança requerido pelo serviço.
5. Existência de Suporte de mercado: Todas as especificações contidas na e-PING contemplam soluções amplamente utilizadas pelo mercado. O objetivo a ser alcançado é a redução dos custos e dos riscos na concepção e produção de serviços nos sistemas de informações governamentais.

2.4.2.2 Infraestrutura Nacional de Dados Abertos

No Brasil, a Infraestrutura Nacional de Dados Abertos (INDA)[19] é a política do governo para dados abertos, formada por um conjunto de padrões, tecnologias, procedimentos e mecanismos de controle necessários para atender às condições de disseminação e compartilhamento de dados e informações públicas, em conformidade com o disposta na e-PING.

Segundo Soares *et al., (2013)* a INDA tem como principais objetivos:

- Proporcionar a busca, o acesso, o reuso e cruzamento dos dados públicos de maneira simples e eficiente.
- Coordenar a padronização na geração, armazenamento, acesso e disseminação dos dados e informações de governo.
- Incentivar a agregação de valor e fomentar a colaboração com o cidadão na implementação de novos serviços.

Baseado nestes princípios, a Cartilha Técnica para Publicação de Dados Abertos

[19] http://www.governoeletronico.gov.br/acoes-e-projetos/Dados-Abertos/inda-infraestrutura-nacional-de-dados-abertos - Acesso em 02.12.2015

no Brasil (2011), sugere a utilização dos seguintes formatos não proprietários para dados abertos:

- JSON (*JavaScript Object Notation*): É um padrão aberto de estruturação de dados baseado em texto e legível por humano.
- XML (*Extensible Markup Language*): É um conjunto de regras para codificar documentos com estrutura hierárquica e em um formato legível por máquina.
- CSV (*Comma-Separated Values*): Valores separados por vírgula, é um formato para armazenamento de dados tabulares em texto.
- ODS (*Open Document Spreadsheet*): É comumente chamado de planilha, similar ao XLS do MS Office Excel, porém aberto, por isso deve ser utilizado em substituição ao XLS.
- RDF (*Resource Description Framework*) é um modelo de dados estruturado em grafos e possui diversos formatos de serialização.

Adicionalmente, é importante ressaltar que nenhum dos padrões citados anteriormente são da área de dados georreferenciados, impossibilitando, por exemplo, o mapeamento de informações utilizando dados com esses padrões.

2.3 Considerações Finais do Capítulo

No desenvolvimento de um *data warehouse* diversos fatores devem ser levados em consideração, sendo que muitas vezes, processos podem resultar em erros, complicações ou até mesmo impedimentos para a execução de outras tarefas. Portanto, os principais desafios se encontram na fase ETL, sendo que as ferramentas de acesso tratam dados provenientes de sistemas heterogêneos com a finalidade de garantir qualidade, integridade e confiabilidade das informações.

Sendo assim, a existência de dados muito complexos torna a implementação de um *data warehouse* consistente um desafio para a equipe, principalmente tratando-se de dados governamentais que por possuírem diversas fontes de dados, torna o processo de integração uma atividade trabalhosa e problemática. Por exemplo, dados provindos de diferentes órgãos poderiam ter sido utilizados com finalidades diferentes, portanto, causariam problemas de sintaxe, formatação e até mesmo questionamentos

sobre o nível de confiabilidade. Além disto, uma quantidade questionável de dados também poderia resultar em problemas de processamento e desempenho.

Por conseguinte, a escolha de ferramentas e definição de parâmetros de aceitabilidade deverá ser definida previamente, assim como o estabelecimento dos requisitos, o que é uma atividade extremamente difícil (Faria, 2006). Tal fato se deve aos usuários que utilizarão o *data warehouse*, assim como suas expectativas. Este público, que inclui gerentes, diretores e analistas, muitas vezes não possuem tempo e disponibilidade para definição de parâmetros e repasse de necessidades, o que poderia causar um descontentamento com o resultado final deste projeto.

Desta maneira, a implantação de um *data warehouse* no governo, possuiria potencial para melhorar a tomada de decisão, assim como facilitar o acesso aos dados públicos e reduzir problemas de baixa produtividade em consequência à demorada de busca de informações e a falta de consistência nos dados devido à diferença das estruturas.

3. Metodologia

A metodologia utilizada neste trabalho será dividida inicialmente em cinco etapas. Em cada uma serão realizadas atividades interdependentes e que se relacionam para a obtenção do objetivo final.

Primeiramente, no referencial teórico, busca-se realizar o levantamento bibliográfico referente ao *data warehouse*, assim como a necessidade de implementação deste ambiente. Para aprofundamento no tema, conceituaremos as aplicações, requisitos, modelos, processos, dificuldades e métodos de análise, além das ferramentas disponíveis no mercado. Ainda na primeira etapa, busca-se explorar o cenário atual das aplicações disponíveis em sistemas governamentais. O contexto de Curitiba também será pesquisado, por meio de dados disponibilizados publicamente pelos institutos IPPUC e IPARDES, além do TCE-PR. Tal estudo ajudará na compreensão da estrutura dos dados, formas de distribuição e problemas de integração.

Na etapa 2, os principais desafios na área de *data warehouse* são analisados, envolvendo dificuldades gerais ao se tratar de ambientes governamentais e possíveis soluções baseadas em estudos relacionados. Esta etapa é importante para o levantamento de questionamentos, como: Quais soluções são aplicáveis a este projeto? Qual a melhor maneira de enfrentar e solucionar dificuldades já conhecidas?

Na etapa 3, será avaliado o subconjunto de dados disponíveis para desenvolvimento deste projeto assim como a identificação dos requisitos. Nesse momento, dados de sistemas reais são coletados, para verificação de suas características. As ferramentas utilizadas para extração de dados serão o SQL Server, em um ambiente Windows 8. Entre as principais características analisadas, estão inclusas: nível de granularidade, formato dos dados, tamanho, frequência, histórico e consistência. Como consequência desta análise, os resultados obtidos serão essenciais para a definição da arquitetura a ser utilizada assim como para o desenvolvimento da etapa mais importante, a construção de um *data warehouse*.

Na etapa 4 e 5, será realizada a construção do *data warehouse*, e posteriormente um estudo de caso sobre o assunto. Neste momento será utilizada a metodologia proposta por Kimbal (1996), que tem como característica possuir os seguintes componentes no DW: (i) fontes de dados, (ii) área de estágio, (iii) área de apresentação de dados e (iv) ferramentas de acesso aos dados. As etapas 4 e 5, podem ser detalhadas com os passos descritos na sequência e ilustrados na Figura 22.

- Modelagem do ambiente

Por possuir alto desempenho e ser de fácil entendimento, nessa fase é estabelecido o tipo de modelagem, além de definir as tabelas de fatos, dimensões e suas estruturas.

- Desenvolvimento do *data warehouse*

Consistirá em implementar o ambiente *data warehousing* planejado, baseando-se nas etapas anteriores. Além disto, neste passo também será avaliado se os requisitos de usuários estão sendo atendidos, assim como o funcionamento do ambiente modelado.

- Implantação

Após o desenvolvimento do *data warehouse*, o ambiente receberá as primeiras cargas de dados, e neste primeiro momento todo o processo será monitorado para garantir o correto funcionamento.

- Estudo de caso

Finalmente, um estudo de caso envolvendo o modelo desenvolvido, aplicado ao contexto do Tribunal de Contas do Estado do Paraná.

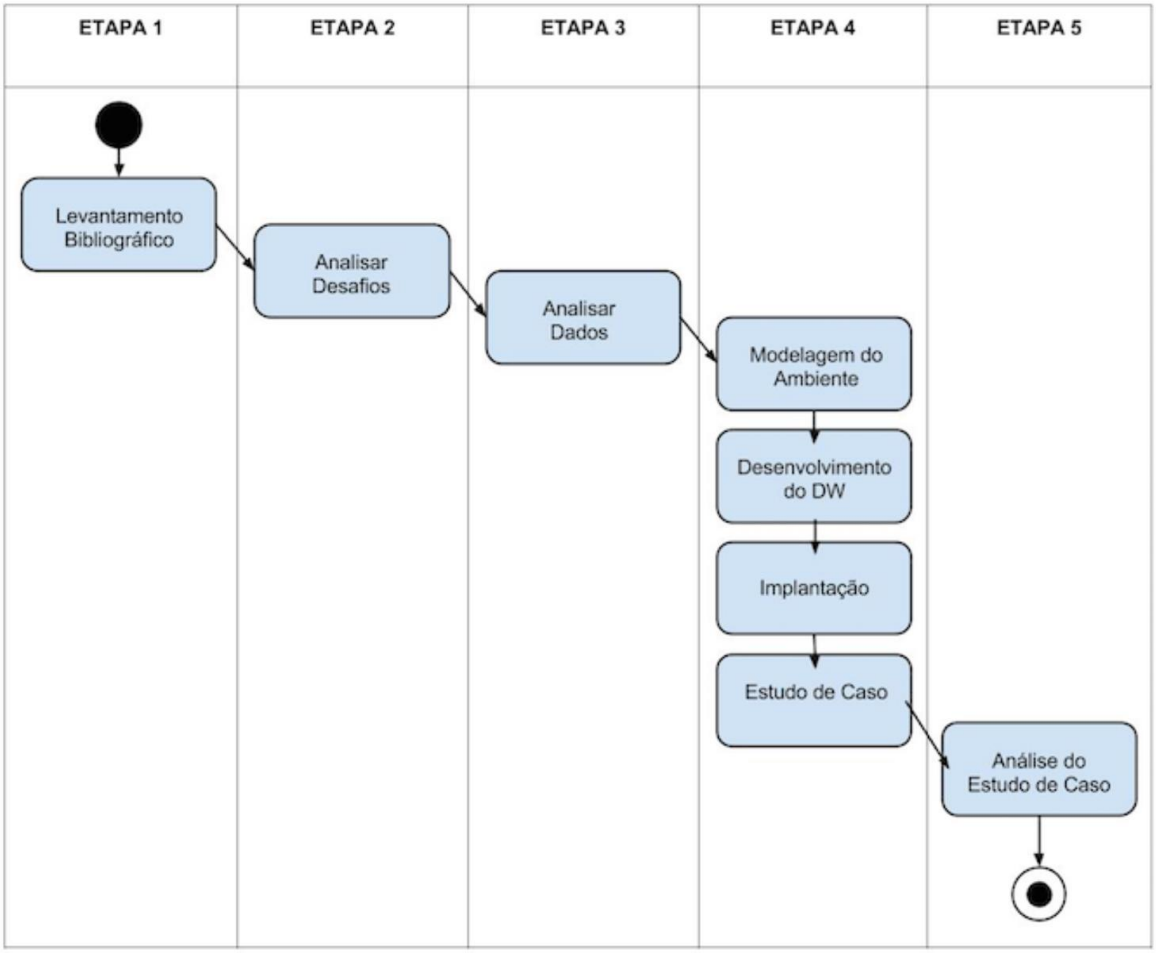


Figura 22: Etapas do projeto. Fonte: Autoria Própria

4. Recursos de *Hardware* e *Software*

Os recursos de hardware e software serão descritos neste capítulo. Os recursos de *hardware* serão expostos na seção 4.1, os recursos de *software* na seção 4.2 e a viabilidade na seção 4.3.

4.1 Hardware

Para análise dos processos e verificação de melhorias do *data warehouse* foi utilizado um servidor dedicado para armazenamento e acesso, com as seguintes características de processamento: Dell Optiplex 380, pequeno porte. O servidor utiliza do sistema operacional Windows 8.

Além deste, foi utilizado também os próprios computadores pessoais já pertencentes aos membros da equipe, que possuem *hardware* compatível com os padrões atuais.

4.2 Software

A plataforma do sistema operacional e a ferramenta de extração de dados já disponíveis para utilização, foram:

• SQL Server 2014 Business Intelligence 64 bits

Essa ferramenta permitiu acessar os dados disponíveis para fins de análise e verificação de padrões. Além disto, o SQL Server proporciona através dos seus serviços de integração, suporte para as tarefas de extração, transação e carregamento (ETL) em um *data warehouse*[20].

• Windows

O sistema operacional utilizado foi o Windows, sabendo que é o sistema atual no Tribunal de Contas do Paraná.

[20] http://www.microsoft.com/pt-br/server-cloud/products/sql-server/ - Acesso em: 19 maio. 2015.

5. Implementação

Neste capítulo é apresentada a parte prática deste projeto, exibindo o desenvolvimento de técnicas, protótipos e modelos utilizados na implementação do estudo de caso referente à construção de um *data warehouse*.

5.1 Caracterização dos Dados

Para que fosse possível a realização deste trabalho, foram utilizadas três tipos de fonte de dados diferentes, sendo que a primeira, foi retirada do sítio do IPARDES, e a segunda e terceira (DEX e Trâmite), foram obtidas com o apoio da equipe de Tecnologia da Informação do Tribunal de Contas do Paraná (TCE-PR). Note, que por motivos de sigilo da informação, alguns dados foram fornecidos de modo anonimizado, como: Nome do gestor, nome da entidade, entre outros. Tais variáveis serão apresentadas no decorrer deste.

Objetivando definir uma estratégia para importação e análise dos dados, primeiramente, foram realizados os procedimentos de caracterização dos dados, viabilizando assim o entendimento dos relacionamentos e informações que tais tabelas agregariam, quando normatizadas e devidamente tratadas, para a criação do da*ta warehouse*. A seguir, é apresentado em detalhes as bases de dados que compõe o escopo aqui previamente definido.

5.1.1 IPARDES

A base de dados utilizada neste trabalho está disponibilizada na página do IPARDES e pode ser acessada por meio do BDEweb[21] que fornece como principais informações dados sobre o Produto Interno Bruto (PIB), Índice de Desenvolvimento Humano (IDH), Índice IPARDES de Desempenho Municipal (IPDM), população, emprego, finanças, agropecuária, entre outros. Para este trabalho, serão utilizados os índices ilustrados na Tabela 3 - Parâmetros IPARDES.

Ou seja, por meio desta base, é possível extrair informações das áreas física, econômica, social, financeira, política e administrativa, disponíveis por municípios, total

[21] http://www.ipardes.pr.gov.br/imp/index.php - Acesso em 02.02.2015

do Estado e para as seguintes agregações: microrregiões geográficas do IBGE, regiões geográficas, regiões metropolitanas e regiões administrativas do Paraná (planejamento, saúde, educação, trabalho, agricultura e comarcas/foros).

Com a finalidade de realizar a análise dos índices, foram selecionados todos os trezentos e noventa e nove municípios, sendo estes pertencentes ao estado do Paraná. Os dados de referência têm como característica serem anuais, sendo assim, para a realização deste projeto foram selecionados os anos de 2008 à 2014 para a base de dados do IPARDES. Adicionalmente, para a extração das informações, depois de selecionados as variáveis, localidades e períodos, optou-se pela única opção disponível para download dos dados, sendo está em formato de arquivo de texto, .CSV. Uma amostra da tabela IPARDES pode ser verificada conforme Tabela 4 – Tabela IPARDES.

IPARDES	
Abastecimento de Água - Ligações	Matrículas no Ensino Fundamental - Rede Municipal
Abastecimento de Água - Ligações Residenciais	Matrículas no Ensino Fundamental - Total
Abastecimento de Água - Unidades Atendidas	Matrículas no Ensino Regular - Rede Municipal
Abastecimento de Água - Unidades Atendidas Residenciais	Matrículas no Ensino Regular - Total
Atendimento de Esgoto - Ligações	Óbitos (CID10) - Total (Mortalidade Geral)
Atendimento de Esgoto - Ligações Residenciais	Óbitos de Menores de 1 ano (CID10) - Total (Mortalidade Infantil)
Atendimento de Esgoto - Unidades Atendidas	População Censitária - Total

Atendimento de Esgoto - Unidades Atendidas Residenciais	População Estimada (IBGE) - Residentes em 01/07
Docentes - Rede Municipal	População Ocupada - Administração Pública, Defesa e Seguridade Social
Docentes - Total	População Ocupada - Agricultura, Pecuária, Produção Florestal, Pesca e Aquicultura
Docentes na Creche - Rede Municipal	População Ocupada - Água, Esgoto, Atividades de Gestão de Resíduos e Descontaminação
Docentes na Creche - Total	População Ocupada - Alojamento e Alimentação
Docentes na Educação Infantil - Rede Municipal	População Ocupada - Artes, Cultura, Esporte e Recreação
Docentes na Educação Infantil - Total	População Ocupada - Atividades Administrativas e Serviços Complementares
Docentes na Pré-Escola - Rede Municipal	População Ocupada - Atividades Financeiras, de Seguros e Serviços Relacionados
Docentes na Pré-Escola - Total	População Ocupada - Atividades Imobiliárias
Docentes no Ensino Fundamental - Rede Municipal	População Ocupada - Atividades mal Especificadas
Docentes no Ensino Fundamental - Total	População Ocupada - Atividades Profissionais, Científicas e Técnicas
Estabelecimentos de Ensino - Rede Municipal	População Ocupada - Comércio Reparação de Veículos Automotores e Motocicletas
Estabelecimentos de Ensino - Total	População Ocupada - Construção
Estabelecimentos de Ensino com Creche - Rede Municipal	População Ocupada - Educação
Estabelecimentos de Ensino com Creche - Total	População Ocupada - Eletricidade e Gás

Estabelecimentos de Ensino Fundamental - Rede Municipal	População Ocupada - Indústrias de Transformação
Estabelecimentos de Ensino Fundamental - Total	População Ocupada - Indústrias Extrativas
Estabelecimentos de Ensino Médio - Rede Municipal	População Ocupada - Informação e Comunicação
Estabelecimentos de Ensino Médio - Total	População Ocupada - Organismos Internacionais e Outras Instituições Extraterritoriais
Estabelecimentos de Ensino Pré-Escolar - Rede Municipal	População Ocupada - Outras Atividades de Serviços
Estabelecimentos de Ensino Pré-Escolar - Total	População Ocupada - Saúde Humana e Serviços Sociais
Grau de Urbanização (%)	População Ocupada - Serviços Domésticos
IDEB - Ensino Fundamental - Anos Iniciais - Rede Municipal	População Ocupada - Total
Índice de Desenvolvimento Humano Municipal (IDH-M)	População Ocupada - Transporte, Armazenagem e Correio
Matrículas na Creche - Rede Municipal	Produto Interno Bruto a Preços Correntes (R$ 1000,00)
Matrículas na Creche - Total	Produto Interno Bruto per Capita (R$ 1,00)
Matrículas na Educação Infantil - Rede Municipal	Taxa de Abandono no Ensino Fundamental - Anos Iniciais (%)
Matrículas na Educação Infantil - Total	Taxa de Analfabetismo de 15 anos ou mais (%)
Matrículas na Pré-Escola - Rede Municipal	Taxa de Aprovação no Ensino Fundamental - Anos Iniciais (%)
Matrículas na Pré-Escola - Total	Taxa de Reprovação no Ensino Fundamental - Anos Iniciais (%)

Tabela 3: Parâmetros utilizados nos dados do IPARDES.

localidade	variavel	ano	valor	cdMunicipio
Iguaraçu	Abastecimento de Água - Ligações	2008	1388	10003
Iguaraçu	Abastecimento de Água - Ligações	2009	NULL	10003
Iguaraçu	Abastecimento de Água - Ligações	2010	1400	10003
Iguaraçu	Abastecimento de Água - Ligações	2012	1609	10003
Iguaraçu	Abastecimento de Água - Ligações	2014	NULL	10003
Iguaraçu	Abastecimento de Água - Ligações Residenciais	2008	NULL	10003
Iguaraçu	Abastecimento de Água - Ligações Residenciais	2009	NULL	10003
Iguaraçu	Abastecimento de Água - Ligações Residenciais	2010	NULL	10003
Iguaraçu	Abastecimento de Água - Ligações Residenciais	2012	NULL	10003
Iguaraçu	Abastecimento de Água - Ligações Residenciais	2014	NULL	10003
Iguaraçu	Abastecimento de Água - Unidades Atendidas	2008	1388	10003
Iguaraçu	Abastecimento de Água - Unidades Atendidas	2009	NULL	10003
Iguaraçu	Abastecimento de Água - Unidades Atendidas	2010	1400	10003
Iguaraçu	Abastecimento de Água - Unidades Atendidas	2012	1609	10003
Iguaraçu	Abastecimento de Água - Unidades Atendidas	2014	NULL	10003
Iguaraçu	Abastecimento de Água - Unidades Atendidas Residenciais	2008	1075	10003
Iguaraçu	Abastecimento de Água - Unidades Atendidas Residenciais	2009	NULL	10003
Iguaraçu	Abastecimento de Água - Unidades Atendidas Residenciais	2010	1400	10003

Tabela 4: Tabela IPARDES.

As variáveis que estruturam a Tabela 5 - IPARDES, utilizadas no processo de associação de registros, são definidas de acordo com a localidade, descrição do indicador e ano de referência. O detalhamento de cada é apresentado conforme apresentado na Tabela 5 – Variáveis IPARDES.

Variável	Descrição
Localidade	Município referente ao indicador
Variável	Descrição do indicador
Ano	Ano referente ao indicador
Valor	Número de pessoas
cdMunicípio	Código do Município

Tabela 5: Variáveis IPARDES.

5.1.2 DEX

A base de dados DEX é originária da diretoria de execuções do TCE, sendo utilizados somente arquivos que já se encontram consolidados, ou seja, não serão realizadas novas publicações contendo alterações nos dados. Nesta base, pode-se obter informações sobre as determinações do TCE, contendo a entidade, motivo da baixa, data, entre outros, conforme pode ser visto na Figura 23.

DEX
Entidade
Gestor
TipoDeterminacao
idTipoDeterminacao
Determinacao
idDeterminacao
Prazo
DecisaoAto
DecisaoUnidade
DecisaoDataRegistro
idEntidade
idGestor
BaixaAto
BaixaMotivo
BaixaUnidade
BaixaDataRegistro
nmMunicipio
cdMunicipio

Figura 23: Tabela DEX. Fonte: Autoria Própria

O TCE é responsável pela fiscalização do uso do dinheiro do público do estado, sendo que quando uma prestação não atende as normas da prestação de contas, são aplicados sansões a gestores, entidades e demais responsáveis (Manual de orientação para o cumprimento de decisões do TCE PR, 10/2015). Assim sendo, a diretoria de execuções do TCE (DEX) é responsável pelo registro, controle e o acompanhamento do cumprimento das sanções e demais determinações, orientando ainda, as entidades públicas acerca das execuções das penalidades aplicadas.

O período dos arquivos, obtidos com o apoio da equipe de Tecnologia da Informação do Tribunal de Contas do Paraná (TCE-PR), compreendem os anos entre 2008 a 2015, todos estes referentes aos municípios que compõe o estado do Paraná (399 municípios). Estes arquivos foram disponibilizados no formato .XLSX e possuem como característica, as datas de registro e baixa das decisões no formato (MM/DD/AA HH:MM). Na Figura 24, é possível visualizar os relacionamentos da tabela DEX.

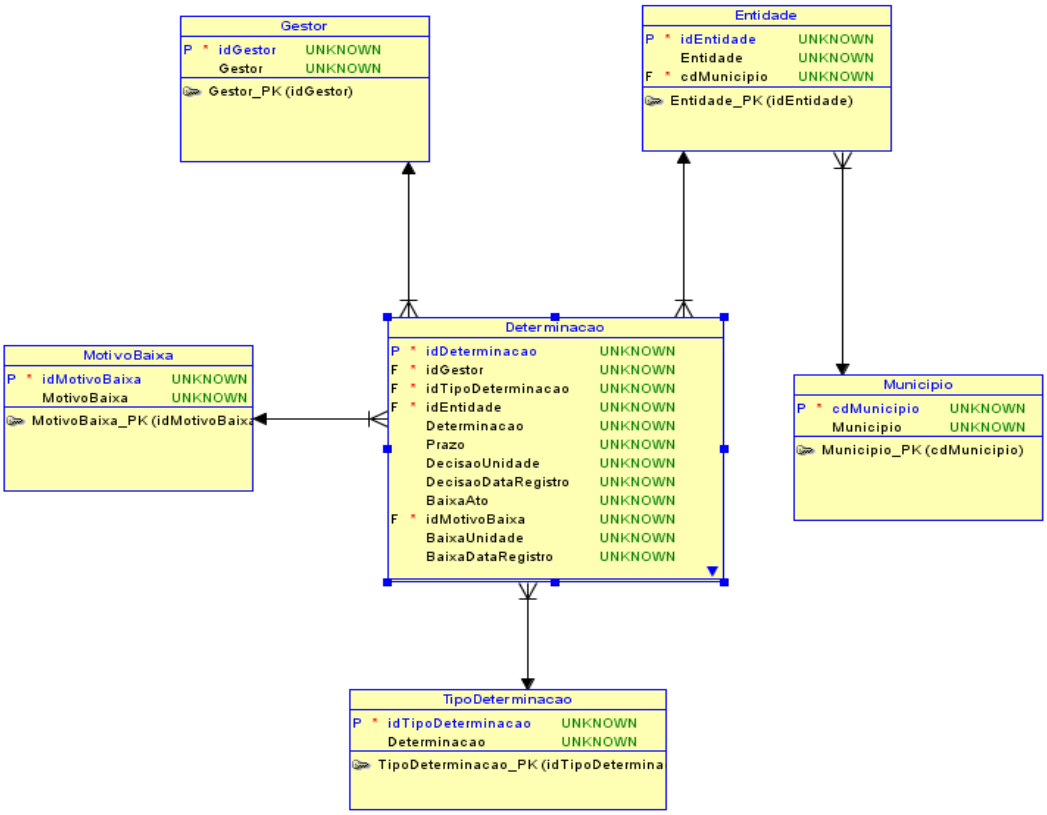


Figura 24: Relacionamento DEX. Fonte: TCEPR

É importante destacar os tipos de determinações que compõe a tabela DEX, sendo estas descritas e apresentadas conforme Tabela 6. Já as variáveis que estruturam esta, utilizadas no processo de associação de registros, são baseadas nas definições de município, entidade, assunto, ato de decisão, relator, localização entre outros. O detalhamento de cada variável é apresentado na Tabela 7. Devido à diversidade de caracteres utilizados na variável “Determinacao”, o que tornou o processo de tratamento de dados muito custoso, optou-se pela remoção do mesmo, sabendo que as informações ali contidas não fazem parte do escopo da análise aqui proposta. Sendo assim, o preenchimento deste campo com o tipo “null”.

DEX	
ABERTURA DE INQUÉRITO ADMINISTRATIVO	ELABORAR PLANO DE AÇÃO
ABERTURA DE PROCEDIMENTO LICITATÓRIO	INGRESSAR COM AÇÃO DE REGRESSO
ABERTURA DE SINDICÂNCIA ADMINISTRATIVA	LEI 8666/93 - ADEQUAÇÃO DO EDITAL À NORMA LEGAL
ABERTURA DE TOMADA DE CONTAS ESPECIAIS	REALIZAÇÃO DE CONCURSO PÚBLICO
ADEQUAÇÃO DA LEGISLAÇÃO MUNICIPAL	REGULARIZAÇÃO DE ADMISSÕES
ALIMENTAÇÃO DO SISTEMA SIM-AP	REGULARIZAÇÃO DE APOSENTADORIAS
ALIMENTAÇÃO DO SISTEMA SIT	REGULARIZAÇÃO DE DÉBITOS PREVIDÊNCIÁRIOS
ALIMENTAÇÃO/AJUSTES NO SISTEMA SIM-AM	REGULARIZAÇÃO DE PENSÕES
APRESENTAÇÃO DE DOCUMENTOS FALTANTES	REGULARIZAÇÃO QUANTO AO RECOLHIMENTO DE TRIBUTOS
APRESENTAÇÃO DE ESCLARECIMENTOS E JUSTIFICATIVAS	REPRESENTAÇÃO DO MP - EXONERAÇÃO DE COMISSIONADO
ATUALIZAÇÃO CADASTRAL JUNTO AO TRIBUNAL DE CONTAS	REPRESENTAÇÃO DO MP - EXTINÇÃO DE CARGO MEDIANTE LEI
COMPROVAÇÃO DE QUITAÇÃO DE DÉBITOS	REPRESENTAÇÃO DO MP - READEQUAÇÃO DA ESTRUTURA DO QUADRO DE CARGOS
CONTAS DO GOVERNADOR - FUNDO DE PREVIDÊNCIA	REVOGAÇÃO DE ATO DE ADMISSÃO DE PESSOAL
CONTROLE INTERNO - COMPROVAÇÃO DE LEGALIDADE DA NOMEAÇÃO	REVOGAÇÃO DE ATO DE APOSENTADORIA
CONTROLE INTERNO - IMPLANTAR O SISTEMA	REVOGAÇÃO DE ATO DE CONCESSÃO DE PENSÃO
CORRETA DESTINAÇÃO DE BEM PÚBLICO	REVOGAÇÃO DE ATO DE REVISÃO DE PROVENTOS
CUMPRIMENTO DE LEGISLAÇÃO EXPRESSA NO ACORDÃO	REVOGAÇÃO DE EDITAL LICITATÓRIO
DESOBEDIÊNCIA À LEI 12.398/98 (PARANAPREVIDÊNCIA)	REVOGAÇÃO DE OUTROS ATOS
DEVOLUÇÃO DE PROCESSOS	SUSTAÇÃO DE ATO IMPUGNADO

Tabela 6: Tipos de determinações DEX. Fonte: Autoria Própria

Variável	Descrição
BaixaAto	Dado anonimizado
BaixaDataRegistro	Data de baixa do registro
BaixaMotivo	Motivo da baixa
BaixaUnidade	Unidade responsável pela baixa
cdMunicipio	Código do município
DecisaoAto	Dado anonimizado
DecisaoDataRegistro	Data de registro da decisão
DecisaoUnidade	Unidade responsável pela decisão
Determinacao	Campo alterado - null
Entidade	Nome do município
Gestor	Dado anonimizado
idDeterminacao	Id da determinação
IdEntidade	Id da entidade
IdGestor	Id do gestor
idTipoDeterminacao	Id do tipo da determinação
nmMunicipio	Nome do município
Prazo	Dias para finalizar uma determinação
TipoDeterminacao	Tipo da determinação

Tabela 7: Variáveis DEX. Fonte: Autoria Própria

5.1.3 Trâmite

A terceira fonte de dados, trâmite, contém arquivos de trâmite de processos com a evolução temporal do estoque. Ou seja, os anos referentes aos dados são caracterizados como anuais se tratando do ano de autuação e mensais ao se referir à data de acompanhamento do processo.

Para fins de análise, foram selecionados seiscentos e vinte e três processos que foram autuados em novembro de 2013 e o acompanhamento mensal deles por treze períodos mensais (de 01/01/2014 até 01/01/2015 – Vide coluna DataRef Figura 25). Como exemplo, pode-se mostrar por meio da Figura 25 as etapas de um processo, sendo que por meio deste, é possível verificar em quais diretorias passou e quanto

tempo ficou. Note que o mesmo processo pode retornar à mesma diretoria diversas vezes, assim como é possível analisar onde o processo ficou mais tempo parado, o processo indicado, por exemplo, esteve durante 8 meses na DICAP, no mês seguinte esteve na GACAC e assim por diante, totalizando um período de 1 ano.

	nrProcesso	Localizacao	DataRef
1	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-01-01
2	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-02-01
3	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-03-01
4	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-04-01
5	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-05-01
6	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-06-01
7	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-07-01
8	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-08-01
9	10B73C2CAC5F0032E38B481302FBF2D7	GACAC	2014-09-01
10	10B73C2CAC5F0032E38B481302FBF2D7	S1C	2014-10-01
11	10B73C2CAC5F0032E38B481302FBF2D7	DICAP	2014-11-01
12	10B73C2CAC5F0032E38B481302FBF2D7	DP	2014-12-01
13	10B73C2CAC5F0032E38B481302FBF2D7	DP	2015-01-01

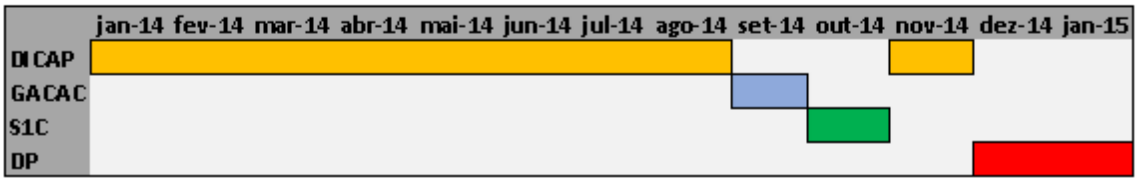


Figura 25: Ciclo de Vida de um Processo. Fonte: Autoria Própria

Todos os processos são referentes aos municípios que compõem o estado do Paraná. Esse arquivo foi disponibilizado para os autores deste trabalho em formato .CSV. A Figura 26 mostra os relacionamentos e atributos da tabela referente aos processos de trâmite.

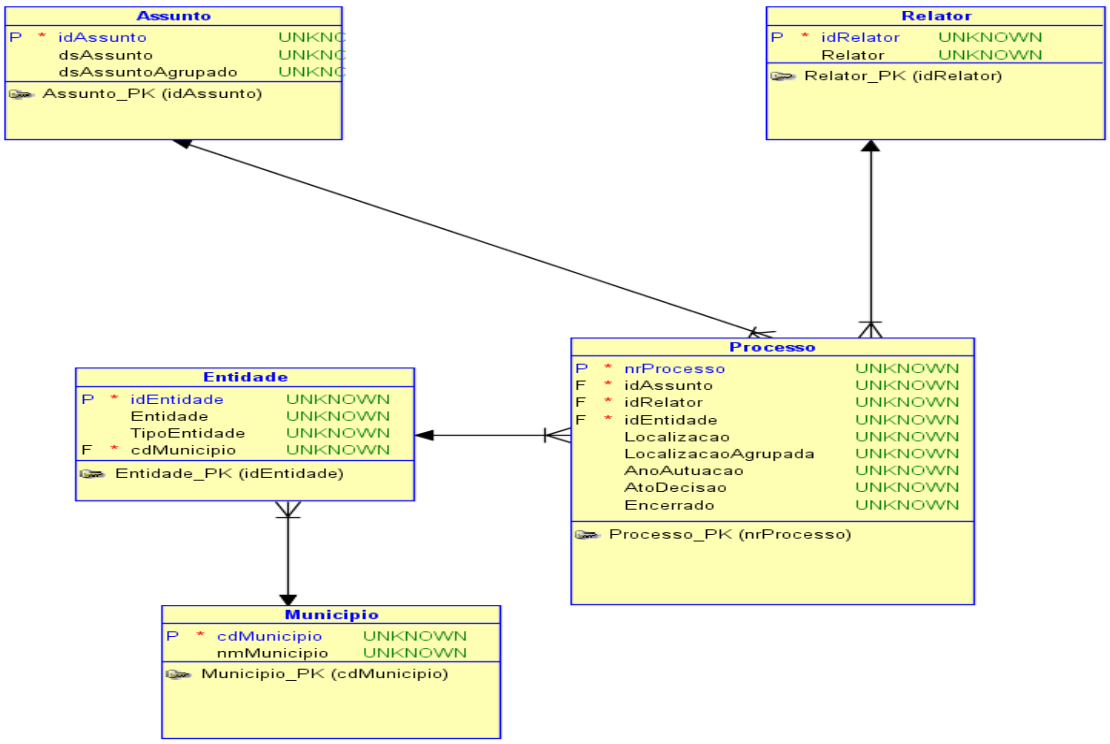


Figura 26: Relacionamentos tabela Trâmite/Processo. Fonte: TCEPR

Os assuntos que foram abordados neste trabalho para a tabela de trâmite/processos, estão descritos conforme Tabela 8. Tais assuntos, também podem ser agrupados, sendo estes representados na Tabela 9. Segundo Art. 330 referente ao regimento interno do TCE-PR, serão autuados como processo os assuntos referidos no Regimento Interno e nas demais Resoluções, mediante Instrução Normativa proposta pela Diretoria-Geral. (Redação dada pela Resolução n° 24/2010). Considera-se então, assunto, a matéria de que trata o processo, consideradas as distintas competências atribuídas por lei ao Tribunal de Contas. (Incluído pela Resolução n° 24/2010).

Trâmite
ADMISSÃO DE PESSOAL
ATO DE INATIVAÇÃO
PEDIDO DE RESCISÃO
PENSÃO
PRESTAÇÃO DE CONTAS ANUAL
PRESTAÇÃO DE CONTAS DE TRANSFERÊNCIA
PRESTAÇÃO DE CONTAS DO PREFEITO MUNICIPAL
PRESTAÇÃO DE CONTAS MUNICIPAL
RECURSO DE AGRAVO
RECURSO DE REVISÃO
RECURSO DE REVISTA
REVISÃO DE PROVENTOS
TOMADA DE CONTAS
TOMADA DE CONTAS EXTRAORDINÁRIA
TOMADA DE CONTAS ORDINÁRIA

Tabela 8: Assuntos – Trâmite.

Assuntos Agrupados
DAT - PRESTAÇÃO DE CONTAS
DCM - PRESTAÇÃO DE CONTAS
RECURSOS
TOMADA DE DCONTAS

Tabela 9: Assuntos Agrupados – Trâmite.

As variáveis que estruturam esta base, utilizadas no processo de associação de registros, são fundamentadas nas definições de município, entidade, assunto, ato de decisão, relator, localização entre outros. O detalhamento de cada variável é apresentado conforme Tabela 10 – Variáveis Trâmite.

Variável	Descrição
AnoAutuação	Ano de autuação
AtoDecisao	Ato de decisão
cdMunicipio	Código do município
dataRef	Data de referência
dsAssuntoAgrupado	Descrição Agrupada
dsAssunto	Descrição do Assunto
dsEncaminhamento	Descrição do encaminhamento Ex: Arquivado
idEntidade	Id da entidade
idAssunto	Id do assunto
LocalizaçãoAgrupada	Localização agrupada
Localização	Localização do processo
Entidade	Nome da entidade anonimizada
nmMunicipio	Nome do município
nrProcesso	Número do processo anominizado
Relator	Relator responsável pelo processo
Encerrado	Status do processo
TipoEntidade	Tipo de entidade pública.

Tabela 10: Variáveis Trâmite.

5.1.4 Municípios

A tabela de municípios, também obtida com o apoio do TCE, traz como principal informação os dados referentes aos municípios do Estado do Paraná (399 registros/cidades). Por meio das variáveis que a compõe, o seu entendimento é fundamental para que, possa ser identificada e conectadas aos seguintes dados.

i) Tabela IPARDES: Localidade

ii) Tabela Trâmite: cdMunicipio

iii) Tabela DEX: cdMunicipio

cdUF	dsUF	cdMesoRegiao	dsMesoRegiao	cdMicroRegiao	dsMicroRegiao	cdMunicipio	dsMunicipio	cdIBGE	sgUF
41	PARANÁ	4103	NORTE CENTRAL PARANAENSE	41006	ASTORGA	10003	IGUARAÇU	411000	PR
41	PARANÁ	4107	SUDOESTE PARANAENSE	41025	CAPANEMA	1002	AMPÉRE	410100	PR
41	PARANÁ	4106	OESTE PARANAENSE	41023	CASCAVEL	10052	IGUATU	411005	PR
41	PARANÁ	4105	CENTRO ORIENTAL PARANAENSE	41019	TELÊMACO BORBA	10078	IMBAÚ	411007	PR
41	PARANÁ	4109	SUDESTE PARANAENSE	41031	PRUDENTÓPOLIS	10102	IMBITUVA	411010	PR
41	PARANÁ	4108	CENTRO-SUL PARANAENSE	41029	GUARAPUAVA	10201	INÁCIO MARTINS	411020	PR
41	PARANÁ	4104	NORTE PIONEIRO PARANAENSE	41015	CORNÉLIO PROCÓPIO	103	ABATIÁ	410010	PR
41	PARANÁ	4101	NOROESTE PARANAENSE	41001	PARANAVAÍ	10300	INAJÁ	411030	PR
41	PARANÁ	4101	NOROESTE PARANAENSE	41003	CIANORTE	10409	INDIANÓPOLIS	411040	PR
41	PARANÁ	4109	SUDESTE PARANAENSE	41031	PRUDENTÓPOLIS	10508	IPIRANGA	411050	PR
41	PARANÁ	4106	OESTE PARANAENSE	41023	CASCAVEL	1051	ANAHY	410105	PR
41	PARANÁ	4101	NOROESTE PARANAENSE	41002	UMUARAMA	10607	IPORÃ	411060	PR
41	PARANÁ	4106	OESTE PARANAENSE	41022	TOLEDO	10656	IRACEMA DO OESTE	411065	PR
41	PARANÁ	4109	SUDESTE PARANAENSE	41032	IRATI	10706	IRATI	411070	PR
41	PARANÁ	4102	CENTRO OCIDENTAL PARANAENSE	41005	CAMPO MOURÃO	10805	IRETAMA	411080	PR
41	PARANÁ	4103	NORTE CENTRAL PARANAENSE	41006	ASTORGA	10904	ITAGUAJÉ	411090	PR
41	PARANÁ	4106	OESTE PARANAENSE	41024	FOZ DO IGUAÇU	10953	ITAIPULÂNDIA	411095	PR
41	PARANÁ	4104	NORTE PIONEIRO PARANAENSE	41015	CORNÉLIO PROCÓPIO	11001	ITAMBARACÁ	411100	PR

Figura 27: Tabela Municípios. Fonte: Autoria Própria

As Variáveis da Tabela Municípios, fornecida pelo TCE PR, utilizadas no processo de associação de registros, podem ser verificadas conforme Tabela 11: Variáveis Municípios.

Variável	Descrição
CdUF	Código referente à Unidade Federativa
dsUF	Descrição da Unidade Federativa
cdMesoRegiao	Código referente à mesorregião
cdMicroRegiao	Código referente à micro região
dsMicroRegiao	Descrição da micro região
cdMunicipio	Código do município
dsMuicipio	Descrição do município
cdIBGE	Código do IBGE
sgUF	Sigla Unidade Federativa

Tabela 11: Variáveis Municípios.

5.2 *Staging Area*

Finalizado o procedimento de entendimento e caracterização dos dados, como parte da estratégia, iniciou-se então o processo de importação e carga dos dados na *Staging Area*. Inicialmente somente uma amostra dos dados foi utilizada para testes por questões de limitação de tempo no escopo definido, assim como pelo fato da falta de conhecimentos específicos sobre os tipos de dados e informações que compõe as tabelas DEX e Trâmite. Na carga inicial, os dados das quatro bases de dados foram carregados em tabelas de dados intermediárias, para que fosse possível realizar a limpeza, transformações e validações. Buscando rapidez nos processos de transformação e carga dos dados, foram criados *filegroups* e índices para otimização

da base (índices em destaque são indicados na Figura 28).

5.2.1 Tabelas da *Staging Area*

As tabelas que compõem a *Staging Area* foram criadas conforme as bases de dados DEX, IPARDES, Munícipios e Trâmite, e estão representadas no formato de tabela única para cada sistema. Sendo assim, por meio da representação na Figura 28, observa-se a composição de cada tabela. Essa abordagem foi definida de acordo com a avaliação do formato dos dados a serem utilizados futuramente na extração dos relatórios.

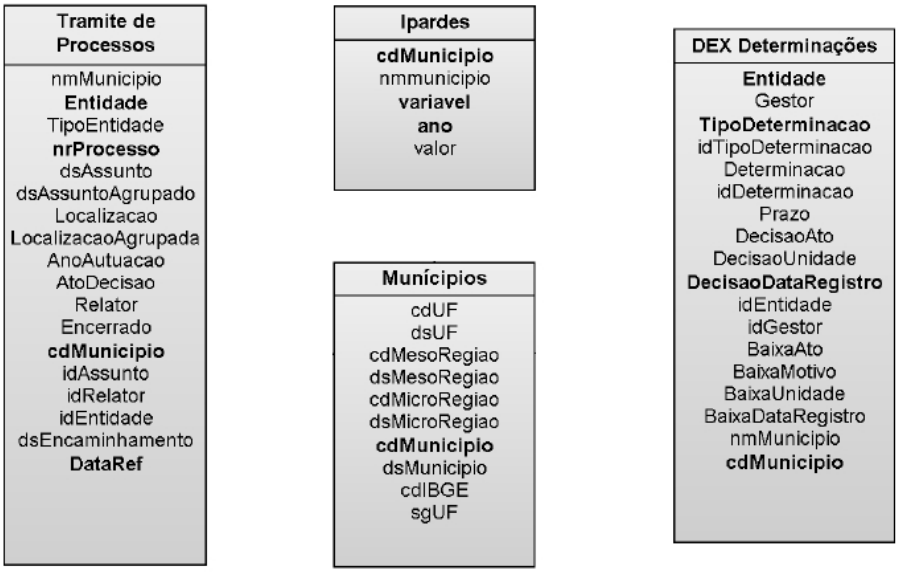


Figura 28: Tabelas da *Staging Area*. Fonte: Autoria Própria

5.2.2 Scripts de limpeza dos dados

Além do procedimento de seleção das variáveis de interesse, o processo de limpeza dos dados foi aplicado por meio de scripts desenvolvidos especificadamente para cada uma das tabelas. A Figura 29 mostra o processo de inserção dos dados (.CSV) na *Staging Area*, utilizando como fonte as tabelas IPARDES, DEX e Trâmite, efetuando assim a limpeza dos dados via scripts (.SH) e inserindo na *Staging Area* por meio de scripts SQL.

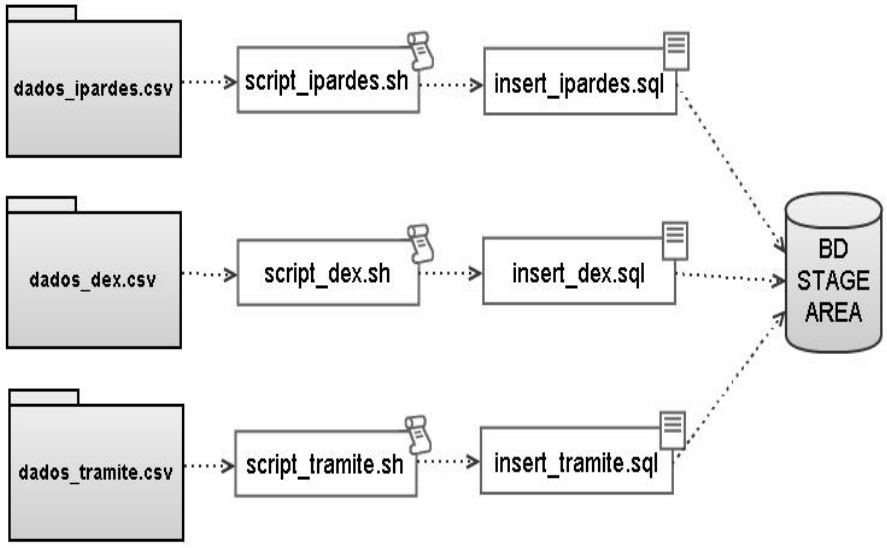


Figura 29: Procedimento de inserção de dados na *Staging Area*. Fonte: Autoria Própria

Visando a exemplificação dos procedimentos efetuados para fins de limpeza dos dados, destaca-se uma amostra do arquivo original (Tabela Trâmite), conforme Figura 30. Analisou-se então, os principais problemas identificados, que quando ocorridos durante a fase de importação dos dados, geram problemas de integração, tornando e caracterizando-os como: (i) Excesso de espaço em branco, (ii) aspas duplas, (iii) separador comum (vírgula) dentro de uma mesma coluna (Ex: “INSTITUTO DE PREVIDÊNCIA, PENSÕES E APOSENTARIA DE SERVIDORES DE ARAPONGAS), entre outros.

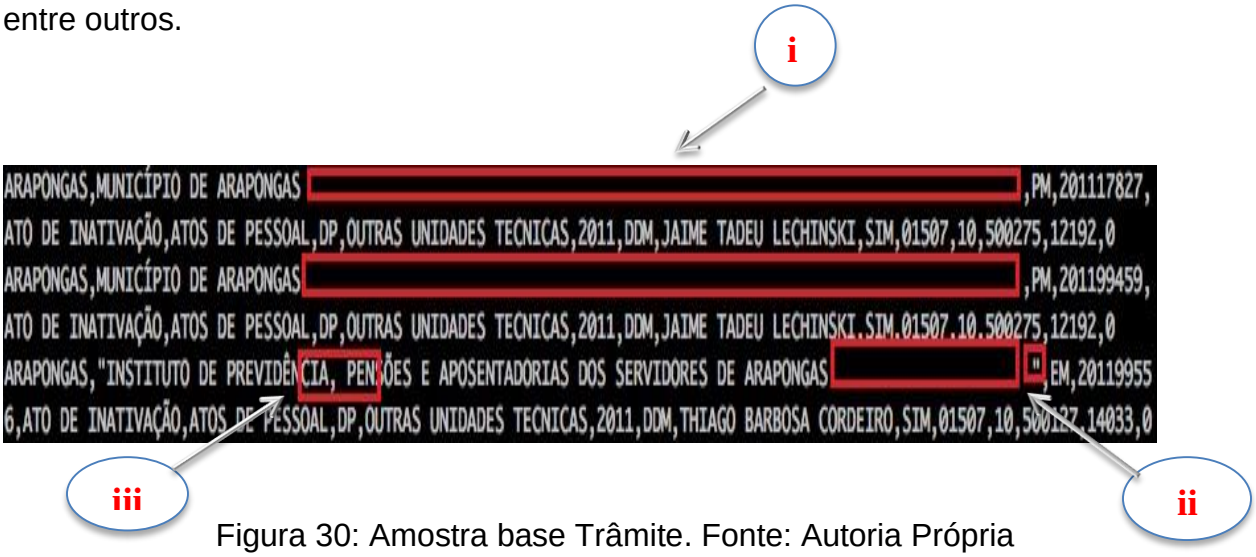


Figura 30: Amostra base Trâmite. Fonte: Autoria Própria

Na Figura 31, é apresentada a solução encontrada, sendo necessário o desenvolvimento de um script, para que assim, fosse viável a realização do processo de tratamento dos dados de maneira mais eficiente. Optou-se pela utilização do Shell Script devido à facilidade de combinação de comandos, podendo utilizá-los um após o outro. Além disso, o mesmo script pode ser reutilizado diversas vezes, gerando assim, um procedimento automático de limpeza dos dados.

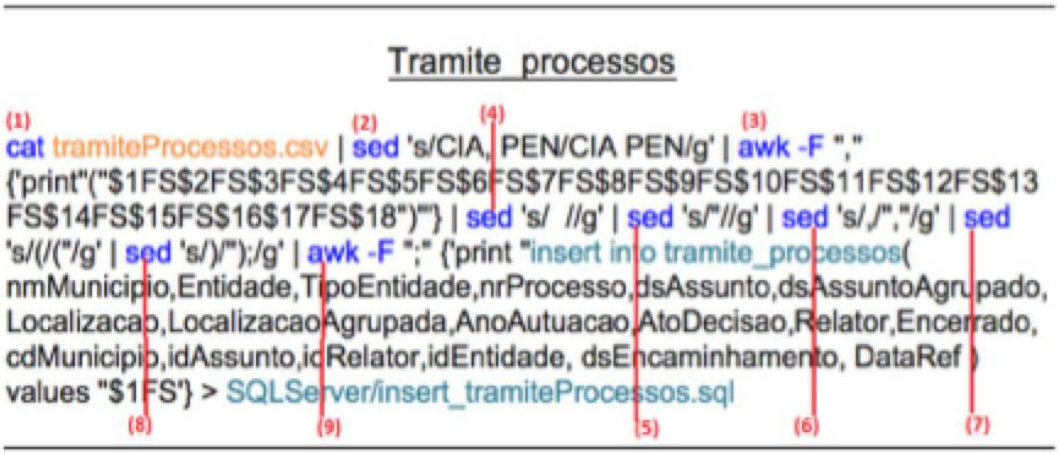


Figura 31: Script de limpeza – Tabela Trâmite. Fonte: Autoria Própria

De maneira resumida, o script faz:

(1) Leitura do arquivo fonte fornecido pelo TCE (Tabela trâmiteProcessos.csv);

(2) Exclusão da vírgula na *string* “CIA, PEN” -> “CIA PEN”, sendo que a vírgula será utilizada como separador de coluna e dessa maneira, geraria problemas para a importação do arquivo.

(3) Definição das colunas, sendo a vírgula caracterizada como separador comum, por meio do comando AWK. O resultado é armazenado entre parênteses, objetivando a utilização do script conforme sintaxe da linguagem SQL.

(4) Limpeza de excesso de espaços em branco;

(5) Limpeza de aspas duplas;

Neste momento, o tratamento dos dados no arquivo fonte é dado como finalizado, portanto, a partir do passo (6), inicia-se o processo de adequação dos comandos para a composição do script de importação de carga .SQL, sendo este denominado SQLServer/insert_trâmiteProcessos.sql;

(6) Substituição de vírgula simples por vírgula entre aspas duplas;
Exemplo: , em “,”

(7) Substituição dos parênteses de abertura em parênteses de abertura mais aspas dupla;
Exemplo: (em (“

(8) Substituição dos parênteses de fechamento em parênteses de fechamento mais aspas dupla;
Exemplo:) em ”)

(9) Integração da parte fixa que compõe a instrução SQL, juntamente com os dados submetidos à limpeza, constituindo assim, o arquivo final de carga.

Já para a base IPARDES, os principais problemas identificados tratam-se de:

(i) Excesso de espaços em branco;

(ii) Inconsistência entre padrões para indicar células vazias, sendo que, algumas apresentam a informação “espaço vazio” e outras traziam um caractere representado por um traço “-”, conforme demonstrado na Figura 32.

O script utilizado para tratamento dos dados será explicado na próxima seção de parametrização.

Figura 32: Amostra base IPARDES. Fonte: Autoria Própria

5.2.3 Parametrização

Devido à complexidade do script e objetivando facilitar os processos de limpeza e carga dos dados no ETL, realizamos a parametrização das variáveis. Todos os scripts seguem o mesmo padrão, portanto, apresenta-se o código desenvolvido para a base de dados IPARDES. Os outros scripts podem ser verificados no apêndice deste trabalho. Desta maneira, também buscou-se facilitar a utilização de outras bases a serem importadas futuramente.

O script, conforme Figura 33, deve ter seu cabeçalho de configuração inicial preparado com as seguintes variáveis:

a. Ano: Ano de referência dos dados;
b. ArquivoEntrada: informar o arquivo de dados que será preparado para carga de dados;
c. ArquivoSaida: informar o arquivo de saída desejado em formato SQL;
d. TabelaBD: Informar a tabela que receberá a carga dos dados.

```
#!/bin/bash

ANO= “2015”
ArquivoEntrada="ipardes.csv"
ArquivoSaida="insert_ipardes.sql"
TabelaBD="TCC.ipardes"

cat $ArquivoEntrada | sed 's/""/null/g' | sed 's/"-"/null/g' | awk -F ";" '{ \
if ($1=="Abatiá") {print "insert into $TabelaBD
(localidade,variavel,ano,valor,cdMunicipio) values ("$1","$2","$ANO,"$3",103"} \

\*
    .
    .   Bloco validação de municípios (IF) vide anexo
    .
*/

}' | sed 's/"\.\.\."/null/g' | sed 's/\.//g' > $ArquivoSaida
```

a. b. c. d.

Figura 33: Script de limpeza parametrizado – base IPARDES. Fonte: Autoria Própria

5.3 Criação do protótipo de Data Warehouse

Concluída a etapa de limpeza dos dados e definição de tabelas no *Staging Area*, por questões de otimização, decidiu-se pela criação de FILEGROUP (FG) no desenvolvimento do protótipo. Foi definida a utilização de índices, buscando mais rapidez em consultas por meio da melhoria de desempenho. Os índices foram implementados juntamente ao FG, e estão disponíveis para consulta no Apêndice C.

As tabelas criadas no DW, seguem a mesma estrutura das tabelas criadas na *Staging Area*, visando otimização, as colunas não utilizadas no relatório foram removidas, conforme mostra a Figura 34.

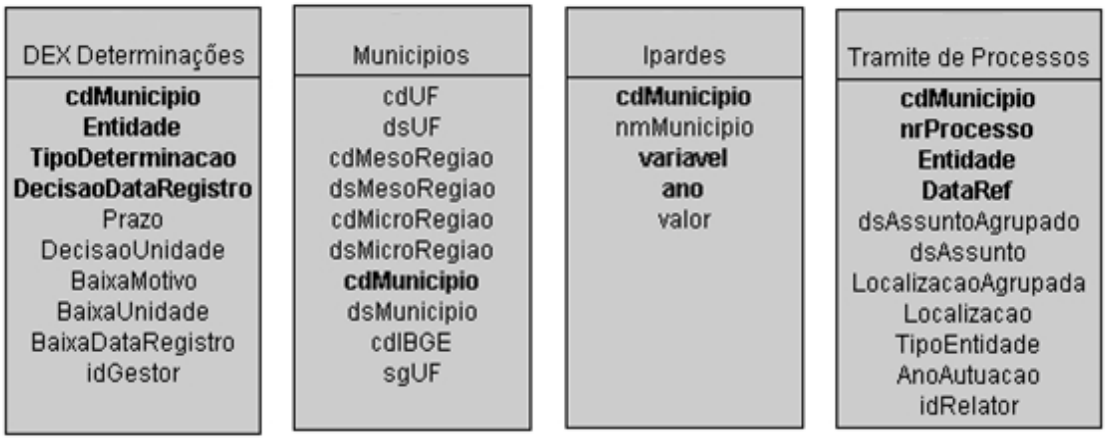


Figura 34: Tabelas criadas no DW. Fonte: Autoria Própria

O armazenamento dos dados foi classificado como "*update*", somente as novas mudanças aplicadas a base de dados é adicionado ao *data warehouse*, não sendo utilizadas operações de exclusão e modificação.

Eventualmente, nesta etapa o modelo estrela poderia ser adaptado e implementado, por possuir mais de uma tabela fato, por meio do conceito de estrela parcial. Neste modelo são utilizadas várias tabelas fato e de dimensão separadas lógica e fisicamente por níveis de sumarização. Sendo assim, os dados seriam caracterizados em níveis de granularidade distintas e existiriam várias estrelas, cada uma representando uma combinação de níveis de agregação em cada dimensão[22]. Para a tabela DEX Determinações, uma possiblidade de partição dos dados em

[22] https://msdn.microsoft.com/pt-br/library/cc518031.aspx Acesso em: 20 dezembro. 2015.

granularidades distintas conforme o modelo estrela poderia ser implementado conforme relações presentes na Figura 24. Para novas agregações, bastaria criar outras tabelas de acordo com a granularidade desejada.

Os dados armazenados no DW estão disponíveis para que usuários e ferramentas analíticas possam acessá-los por meio de queries, conforme apresentado na próxima seção – Etapas de testes e resultados preliminares.

5.3.1 Etapas de testes e resultados preliminares

Nesta etapa, os resultados preliminares são ilustrados e exemplificados, utilizando os dados disponíveis no *data warehouse* para produção de relatórios, agregando informações dentro de suas possibilidades, tornando-as o mais interessante em nível de gestão pública.

Inicialmente, o TCE disponibilizou um modelo de relatório para resultados de consultas de dados por munícipio, utilizando como base, informações das tabelas trâmite de processos e DEX, respectivamente. O relatório apresenta um painel de informações sumarizadas conforme apresentado na Figura 35, Figura 36 e Figura 37, que ilustram o escopo de dados deste modelo.

Tipos de Assunto

TIPOS DE ASSUNTO	Nº
ATOS DE PESSOAL	214
ADMISSÃO DE PESSOAL	43
ATO DE INATIVAÇÃO	84
PENSÃO	31
REVISÃO DE PENSÃO	1
REVISÃO DE PROVENTOS	55
DAT - PRESTAÇÕES DE CONTAS	33
PC DE TRANSFERÊNCIA	33
DCM - PRESTAÇÕES DE CONTAS	22
PC ANUAL	19
PC DO PREFEITO MUNICIPAL	2
PC MUNICIPAL	1
RECURSOS	3
RECURSO DE REVISTA	3
TOMADAS DE CONTAS	2
TC EXTRAORDINÁRIA	2
TOTAL	274

Localização dos Processos

GABINETES E	PM	CM	EM	TOTAL
DAT	35			35
DCM	2	1	18	21
DICAP	138	4	6	148
OUTRAS UN TECNICAS	38			38
DP	37			37
S2C	1			1
GACAC	2		1	3
GASRVF	3			3
GATBC	5			5
GCAML	6			6
GCDA	1		1	2
OUTROS GABINETES			1	1
SMPJTC	3			3
SOBRESTADO	8		1	9
TOTAL	241	5	28	274

Figura 35: Modelo de relatório 1 – Trâmite de Processos. Fonte: TCEPR

Entidades

ENTIDADES	Nº
MUNICÍPIO DE ARAUCÁRIA	241
CÂMARA MUNICIPAL DE ARAUCARIA	5
COMPANHIA DE DESENVOLVIMENTO DE ARAUCÁRIA	5
COMPANHIA MUNICIPAL DE HABITAÇÃO DE ARAUCÁRIA	11
COMPANHIA MUNICIPAL DE TRANSPORTE COLETIVO DE ARAUCÁRIA	6
FUNDO DE PREVIDENCIA MUNICIPAL DE ARAUCARIA	4
SECRETARIA MUNICIPAL DE EDUCAÇÃO DE ARAUCARIA	1
SECRETARIA MUNICIPAL DE SAUDE DE ARAUCARIA	1
TOTAL	274

Ano de Autação

TIPOS DE ASSUNTO	2010	2011	2012	2013	2014	2015	≤ 2009	TOTAL
ATOS DE PESSOAL	3	4	32	129	39	5	2	214
ADMISSÃO DE PESSOAL	3		17	10	11		2	43
ATO DE INATIVAÇÃO		3	11	55	15			84
PENSÃO		1	4	11	12	3		31
REVISÃO DE PENSÃO						1		1
REVISÃO DE PROVENTOS				53	1	1		55
DAT - PRESTAÇÕES DE CONTAS				32		1		33
PC DE TRANSFERÊNCIA				32		1		33
DCM - PRESTAÇÕES DE CONTAS		2	3	3	7	6	1	22
PC ANUAL		2	3	3	6	5		19
PC DO PREFEITO MUNICIPAL					1	1		2
PC MUNICIPAL							1	1
RECURSOS		1			1	1		3
RECURSO DE REVISTA		1			1	1		3
TOMADAS DE CONTAS				2				2
TC EXTRAORDINÁRIA				2				2
TOTAL	3	7	35	166	47	13	3	274

Figura 36: Modelo de relatório 2 – Trâmite de Processos. Fonte: TCEPR

DECISÕES TRANSITADAS EM JULGADO

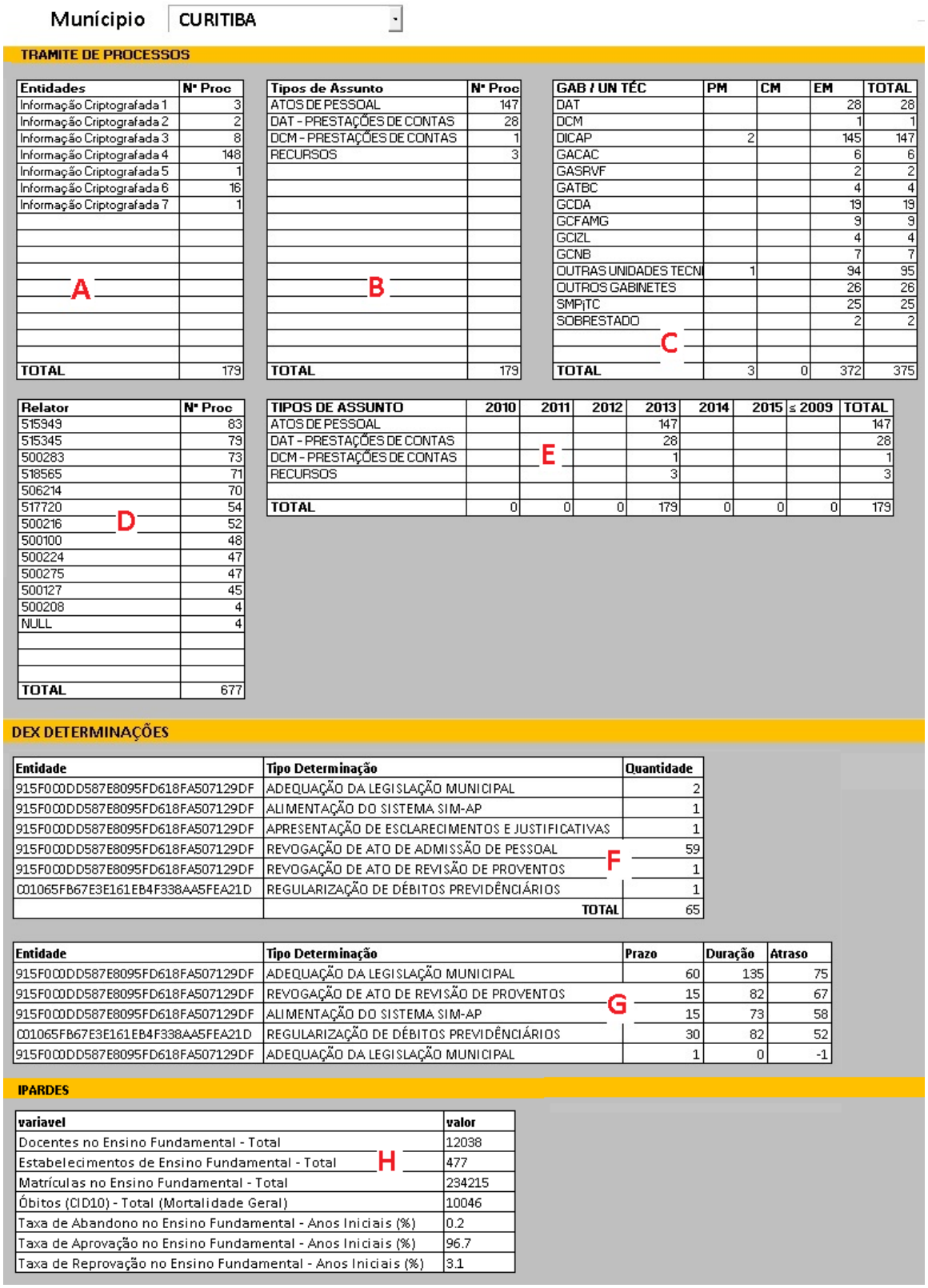
Determinações

ENTIDADE / TIPO DE DETERMINAÇÃO	QUANTIDADE
MUNICÍPIO DE ARAUCÁRIA	2
APRESENTAÇÃO DE ESCLARECIMENTOS E JUSTIFICATIVAS	1
REPRESENTAÇÃO DO MP - EXONERAÇÃO DE COMISSIONADO	1
FUNDO DE PREVIDENCIA MUNICIPAL DE ARAUCARIA	2
APRESENTAÇÃO DE DOCUMENTOS FALTANTES	1
REVOGAÇÃO DE ATO DE APOSENTADORIA	1
COMPANHIA MUNICIPAL DE HABITAÇÃO DE ARAUCÁRIA	1
CUMPRIMENTO DE LEGISLAÇÃO EXPRESSA NO ACORDÃO	1
COMPANHIA MUNICIPAL DE TRANSPORTE COLETIVO DE ARAUCÁRIA	1
CONTAS DO GOVERNADOR - FUNDO DE PREVIDÊNCIA	1
TOTAL	6

Figura 37: Modelo de relatório - DEX. Fonte: TCEPR

Baseado no protótipo fornecido, a Figura 38 propõe um novo modelo de relatório que pode ser extraído por meio das bases que compõe o escopo deste projeto, conforme os itens sinalizados na imagem. Na sequência discute-se cada tabela e suas respectivas consultas na base de dados, utilizando como exemplo, a simulação para o munícipio de Curitiba. Dentre as novas consultas, são incluídas, por exemplo:

- Relatório do número de processos por Relator, proporcionando visualizar a divisão de responsabilidades atribuídas a cada responsável de modo individual.
- Relatório do tipo de determinações por entidade com seu respectivo prazo e duração média (dias) para conclusão. Apresenta-se também, o número médio de dias em atraso;
- Também foram selecionadas variáveis relevantes ao município em foco no relatório que compõe o escopo de dados da tabela IPARDES.



Munícipio CURITIBA

TRAMITE DE PROCESSOS

Entidades	Nº Proc
Informação Criptografada 1	3
Informação Criptografada 2	2
Informação Criptografada 3	8
Informação Criptografada 4	148
Informação Criptografada 5	1
Informação Criptografada 6	16
Informação Criptografada 7	1
TOTAL	179

A

Tipos de Assunto	Nº Proc
ATOS DE PESSOAL	147
DAT - PRESTAÇÕES DE CONTAS	28
DCM - PRESTAÇÕES DE CONTAS	1
RECURSOS	3
TOTAL	179

B

GAB / UN TÉC	PM	CM	EM	TOTAL
DAT			28	28
DCM			1	1
DICAP	2		145	147
GACAC			6	6
GASRVF			2	2
GATBC			4	4
GCDA			19	19
GCFAMG			9	9
GCIZL			4	4
GCNB			7	7
OUTRAS UNIDADES TECNI	1		94	95
OUTROS GABINETES			26	26
SMPjTC			25	25
SOBRESTADO			2	2
TOTAL	3	0	372	375

C

Relator	Nº Proc
515349	83
515345	79
500283	73
518565	71
506214	70
517720	54
500216	52
500100	48
500224	47
500275	47
500127	45
500208	4
NULL	4
TOTAL	677

D

TIPOS DE ASSUNTO	2010	2011	2012	2013	2014	2015	≤ 2009	TOTAL
ATOS DE PESSOAL				147				147
DAT - PRESTAÇÕES DE CONTAS				28				28
DCM - PRESTAÇÕES DE CONTAS				1				1
RECURSOS				3				3
TOTAL	0	0	0	179	0	0	0	179

E

DEX DETERMINAÇÕES

Entidade	Tipo Determinação	Quantidade
915F00DD587E8095FD618FA507129DF	ADEQUAÇÃO DA LEGISLAÇÃO MUNICIPAL	2
915F00DD587E8095FD618FA507129DF	ALIMENTAÇÃO DO SISTEMA SIM-AP	1
915F00DD587E8095FD618FA507129DF	APRESENTAÇÃO DE ESCLARECIMENTOS E JUSTIFICATIVAS	1
915F00DD587E8095FD618FA507129DF	REVOGAÇÃO DE ATO DE ADMISSÃO DE PESSOAL	59
915F00DD587E8095FD618FA507129DF	REVOGAÇÃO DE ATO DE REVISÃO DE PROVENTOS	1
C01065FB67E3E161EB4F338AA5FEA21D	REGULARIZAÇÃO DE DÉBITOS PREVIDÊNCIÁRIOS	1
	TOTAL	65

F

Entidade	Tipo Determinação	Prazo	Duração	Atraso
915F00DD587E8095FD618FA507129DF	ADEQUAÇÃO DA LEGISLAÇÃO MUNICIPAL	60	135	75
915F00DD587E8095FD618FA507129DF	REVOGAÇÃO DE ATO DE REVISÃO DE PROVENTOS	15	82	67
915F00DD587E8095FD618FA507129DF	ALIMENTAÇÃO DO SISTEMA SIM-AP	15	73	58
C01065FB67E3E161EB4F338AA5FEA21D	REGULARIZAÇÃO DE DÉBITOS PREVIDÊNCIÁRIOS	30	82	52
915F00DD587E8095FD618FA507129DF	ADEQUAÇÃO DA LEGISLAÇÃO MUNICIPAL	1	0	-1

G

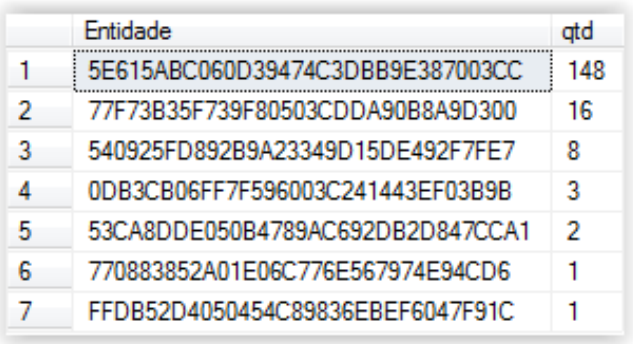
IPARDES

variavel	valor
Docentes no Ensino Fundamental - Total	12038
Estabelecimentos de Ensino Fundamental - Total	477
Matrículas no Ensino Fundamental - Total	234215
Óbitos (CID10) - Total (Mortalidade Geral)	10046
Taxa de Abandono no Ensino Fundamental - Anos Iniciais (%)	0.2
Taxa de Aprovação no Ensino Fundamental - Anos Iniciais (%)	96.7
Taxa de Reprovação no Ensino Fundamental - Anos Iniciais (%)	3.1

H

Figura 38: Novo modelo de relatório. Fonte: Autoria Própria

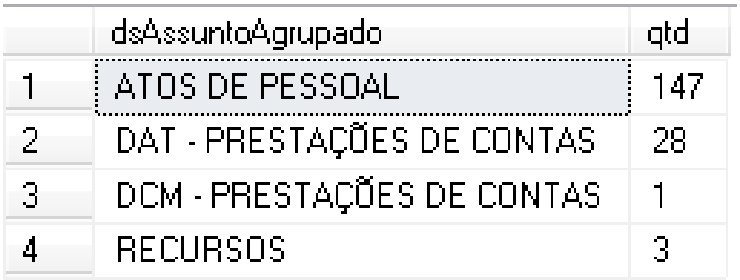
- **Tabela A**: Apresenta uma sumarização com número de processos em cada entidade disponível no munícipio com dados extraídos da tabela trâmite de processos conforme resultado ilustrado na Figura 39.
    - **SQL:** select Entidade, count(distinct nrProcesso) as qtd from DW_TCE.trâmite_processos where cdMunicipio='6902' group by entidade order by 2;



	Entidade	qtd
1	5E615ABC060D39474C3DBB9E387003CC	148
2	77F73B35F739F80503CDDA90B8A9D300	16
3	540925FD892B9A23349D15DE492F7FE7	8
4	0DB3CB06FF7F596003C241443EF03B9B	3
5	53CA8DDE050B4789AC692DB2D847CCA1	2
6	770883852A01E06C776E567974E94CD6	1
7	FFDB52D4050454C89836EBEF6047F91C	1

Figura 39: Consulta no DW – Tabela A. Fonte: Autoria Própria

- **Tabela B**: Apresenta uma sumarização com número de processos por assunto agrupado com dados extraídos da tabela trâmite de processos conforme resultado ilustrado na Figura 40.
    - **SQL:** select dsAssuntoAgrupado, count(distinct nrProcesso) as qtd from DW_TCE.trâmite_processos where cdMunicipio='6902' group by dsAssuntoAgrupado order by 2;



	dsAssuntoAgrupado	qtd
1	ATOS DE PESSOAL	147
2	DAT - PRESTAÇÕES DE CONTAS	28
3	DCM - PRESTAÇÕES DE CONTAS	1
4	RECURSOS	3

Figura 40: Consulta no DW – Tabela B. Fonte: Autoria Própria

- **Tabela C**: Apresenta uma sumarização com número de processos por localização agrupada e combinados por tipo de entidade, dados extraídos da tabela trâmite de processos conforme resultado ilustrado na Figura 41.
    - **SQL:** select Localizacaoagrupada, tipoentidade, count(distinct nrProcesso) as qtd from DW_TCE.trâmite_processos where cdMunicipio='6902' group by Localizacaoagrupada, tipoentidade order by 3 DESC;

	Localizacaoagrupada	tipoentidade	qtd
1	DICAP	EM	145
2	OUTRAS UNIDADES TECNICAS	EM	94
3	DAT	EM	28
4	OUTROS GABINETES	EM	26
5	SMPjTC	EM	25
6	GCDA	EM	19
7	GCFAMG	EM	9
8	GCNB	EM	7
9	GACAC	EM	6
10	GATBC	EM	4
11	GCIZL	EM	4
12	GASRVF	EM	2
13	SOBRESTADO	EM	2
14	DICAP	PM	2
15	OUTRAS UNIDADES TECNICAS	PM	1
16	DCM	EM	1

Figura 41: Consulta no DW – Tabela C. Fonte: Autoria Própria

- **Tabela D**: Apresenta uma sumarização com número de processos por relator, dados extraídos da tabela trâmite de processos conforme resultado ilustrado na Figura 42.
    - **SQL:** select idrelator, count (distinct nrProcesso) as qtd from DW_TCE.trâmite_processos group by idrelator order by 2 DESC;

	idrelator	qtd
1	515949	83
2	515345	79
3	500283	73
4	518565	71
5	506214	70
6	517720	54
7	500216	52
8	500100	48
9	500224	47
10	500275	47
11	500127	45
12	500208	4
13	NULL	4

Figura 42: Consulta no DW – Tabela D. Fonte: Autoria Própria

- **Tabela E**: Apresenta uma sumarização com número de processos por ano de atuação e combinado por assunto agrupado, dados extraídos da tabela trâmite de processos conforme resultado ilustrado na Figura 43.
    - **SQL:** select dsAssuntoAgrupado, anoautuacao, count(distinct nrProcesso) as qtd from DW_TCE.trâmite_processos where cdMunicipio='6902' group by dsAssuntoAgrupado, anoautuacao order by 3 DESC;

	dsAssuntoAgrupado	anoautuacao	qtd
1	ATOS DE PESSOAL	2013	147
2	DAT - PRESTAÇÕES DE CONTAS	2013	28
3	DCM - PRESTAÇÕES DE CONTAS	2013	1
4	RECURSOS	2013	3

Figura 43: Consulta no DW – Tabela E. Fonte: Autoria Própria

- **Tabela F**: Apresenta uma sumarização com número de determinações de cada entidade agrupado pelo tipo da determinação, dados extraídos da tabela DEX conforme resultado ilustrado na Figura 44.
    - **SQL:** select entidade, TipoDeterminacao, count(*) as qtd from DW_TCE.dex_determinacoes where cdMunicipio = '01101' group by entidade, TipoDeterminacao order by 3 DESC;

	entidade	TipoDeterminacao	qtd
1	915F0C0DD587E8095FD618FA507129DF	REVOGAÇÃO DE ATO DE ADMISSÃO DE PESSOAL	59
2	915F0C0DD587E8095FD618FA507129DF	ADEQUAÇÃO DA LEGISLAÇÃO MUNICIPAL	2
3	915F0C0DD587E8095FD618FA507129DF	ALIMENTAÇÃO DO SISTEMA SIM-AP	1
4	915F0C0DD587E8095FD618FA507129DF	APRESENTAÇÃO DE ESCLARECIMENTOS E JUSTIFICATIVAS	1
5	C01065FB67E3E161EB4F338AA5FEA21D	REGULARIZAÇÃO DE DÉBITOS PREVIDÊNCIÁRIOS	1
6	915F0C0DD587E8095FD618FA507129DF	REVOGAÇÃO DE ATO DE REVISÃO DE PROVENTOS	1

Figura 44: Consulta no DW – Tabela F. Fonte: Autoria Própria

- **Tabela G**: Apresenta uma sumarização temporal sobre o prazo estipulado para cada tipo de determinação, a média do tempo de baixa destas determinações e o tempo médio de atraso para cada entidade em um tipo de determinação, dados extraídos da tabela DEX conforme resultado ilustrado na Figura 45.
    - **SQL:** select entidade, TipoDeterminacao, Prazo, AVG(DATEDIFF(day,DecisaoDataRegistro,BaixaDataRegistro)) as DuracaoDias,(CONVERT(INT,AVG(DATEDIFF(day,DecisaoDataRegistro,BaixaDataRegistro))) - CONVERT(INT, Prazo)) as AtrasoDias from DW_TCE.dex_determinacoes where BaixaDataRegistro is not null AND cdmunicipio='01101' group by entidade, TipoDeterminacao, Prazo order by AtrasoDias DESC;

	entidade	TipoDeterminacao	Prazo	DuracaoDias	AtrasoDias
1	915F0C0DD587E8095FD618FA507129DF	ADEQUAÇÃO DA LEGISLAÇÃO MUNICIPAL	60	135	75
2	915F0C0DD587E8095FD618FA507129DF	REVOGAÇÃO DE ATO DE REVISÃO DE PROVENTOS	15	82	67
3	915F0C0DD587E8095FD618FA507129DF	ALIMENTAÇÃO DO SISTEMA SIM-AP	15	73	58
4	C01065FB67E3E161EB4F338AA5FEA21D	REGULARIZAÇÃO DE DÉBITOS PREVIDÊNCIÁRIOS	30	82	52
5	915F0C0DD587E8095FD618FA507129DF	ADEQUAÇÃO DA LEGISLAÇÃO MUNICIPAL	1	0	-1

Figura 45: Consulta no DW – Tabela G. Fonte: Autoria Própria

- **Tabela H**: Apresenta uma sumarização de algumas variáveis disponíveis nos dados do IPARDES para o ano de 2012 conforme resultado ilustrado na figura 46.
    - **SQL:** select variavel, valor from TCC.ipardesnew where cdMunicipio='6902' AND ano='2012' AND (variavel='Docentes no Ensino Fundamental - Total' OR variavel='Estabelecimentos de Ensino Fundamental - Total' OR variavel='Matrículas no Ensino Fundamental - Total' OR variavel='Óbitos (CID10) - Total (Mortalidade Geral)' OR variavel='Taxa de Abandono no Ensino Fundamental - Anos Iniciais (%)' OR variavel='Taxa de Aprovação no Ensino Fundamental - Anos Iniciais (%)' OR variavel='Taxa de Reprovação no Ensino Fundamental - Anos Iniciais (%)');

	variavel	valor
1	Docentes no Ensino Fundamental - Total	12038
2	Estabelecimentos de Ensino Fundamental - Total	477
3	Matrículas no Ensino Fundamental - Total	234215
4	Óbitos (CID10) - Total (Mortalidade Geral)	10046
5	Taxa de Abandono no Ensino Fundamental - Anos Iniciais (%)	0.2
6	Taxa de Aprovação no Ensino Fundamental - Anos Iniciais (%)	96.7
7	Taxa de Reprovação no Ensino Fundamental - Anos Iniciais (%)	3.1

Figura 46: Consulta no DW – Tabela H. Fonte: Autoria Própria

De modo geral, o relatório apresenta uma sumarização de informações por munícipio, o aprofundamento em informações com maior agregação e relacionamento está limitado ao escopo de dados utilizado neste trabalho. Destaca-se no relatório ilustrado na Figura 38, a tabela sinalizada G, pela utilização de funções temporais, fornecendo informações interessantes para gestão das entidades envolvidas.

5.4 Análise

O uso da modelagem multidimensional pode trazer diversos benefícios relacionados ao desempenho de consultas, por meio da utilização tabelas fatos e dimensões. Entretanto, a utilização de dados de diferentes órgãos governamentais tornou-se um desafio à medida em que o processo de limpeza de dados tornou-se muito extenso, sendo necessário desempenhar uma grande quantidade de tempo para que fosse possível realizar a integração dos dados de maneira confiável.

Dito isto, é importante de ressaltar os pontos em que os autores concentraram a maior parte dos recursos, mais especificadamente, tempo, sendo eles:

1. Qualidade dos dados comprometida, na qual houve um grande esforço no desenvolvimento de scripts que realizassem as atividades de inspeção da integridade das bases de dados, assim como a limpeza e tratamento destes dados.

2. Automatização do processo de limpeza e inserção no SQL Server utilizando Shell Script, permitindo que novos dados sejam carregados no DW sem a necessidade de desenvolver novos métodos.

3. Caracterização dos tipos dos dados, sendo que, para uma análise completa seria necessário aprofundar-se no entendimento dos processos de negócio, o que ultrapassaria o tempo disponível para a implantação do DW.

4. A definição do nível de agregação temporal mostrou-se trabalhosa, pelo fato das bases de dados possuírem características diferenciadas. Desta maneira, a utilização da agregação mensal, foi definida conforme o entendimento dos dados e a análise das formas de integração entre as bases, de forma a tornar possível a extração de informações íntegras e completas.

Visto que, devido à restrição da quantidade de dados utilizados neste trabalho, não foi possível realizar a medição de ganho em relação ao desempenho em relação à carga no DW, citados na utilização de índices e *filegroups*. Ainda assim, o modelo contempla tais funções, que poderão impactar positivamente na utilização do mesmo para novas cargas futuras.

Adicionalmente, por questões de melhoria de desempenho, optou-se pela não utilização das colunas consideradas irrelevantes, por não proporcionarem informações úteis ao relatório. Entretanto, estas colunas poderiam ser adicionadas facilmente nas devidas tabelas, ampliando assim, o escopo de informações disponíveis ao relatório.

6. Conclusão

A dificuldade de comparação de informações e a necessidade de se extrair e analisar informações úteis provindas de fontes heterogêneas, que juntas, trazem informações que apoiam a tomada de decisão, foi a questão motivadora deste trabalho. Esta questão conduziu a proposta de criação de um ambiente de *data warehouse* para a extração, limpeza e análise dos dados, a partir das bases de dados do TCE-PR e IPARDES.

A partir deste pressuposto, foi definido, analisado e implementado um modelo de solução por meio de recursos e ferramentas previamente definidos na análise de viabilidade deste projeto.

Sendo assim, um conjunto de objetivos em específico foi atendido por meio do estudo de caso, sendo estes:

1. Revisão literária sobre:
    a. Criação de *data warehouse;*
    b. Dados governamentais;
    c. Padrões atuais.

2. Definição e a criação de um protótipo de *data warehouse,* assim como a modelagem do ambiente, desenvolvimento e implantação. Por meio de um subconjunto de dados fornecidos pelo TCE-PR esse protótipo foi carregado com dados da Tabela DEX e da Tabela Trâmite, além de dados do público obtidos no site do IPARDES.

3. Desenvolvimento de scripts para limpeza e importação dos dados, permitindo assim a carga dos dados na *Staging Area* e no DW, aprimorado desta maneira, a qualidade e estrutura dos dados. Vale ressaltar que esse item em específico foi de extrema importância no processo de integração dos dados, sendo a maior parte do tempo gasto nessa etapa.

4. Foram identificados padrões e inconsistências dos dados provindos de bases heterogêneas, e por meio do processo de limpeza e transformação, foram propostas técnicas de padronização. Também foram expostos processos de melhoria de desempenho (índices e *filegroups*), por meio da implementação do modelo multidimensional proposto, sendo este, baseado em tabelas de fatos.

5. O modelo estrela pode ser aplicado à diversos níveis de granularidade de acordo com a necessidade, para todas as tabelas de fato apresentadas, conforme demonstrado no trabalho.

6. Adaptação dos resultados obtidos em forma de relatórios, de maneira que possibilitem a análise e comparação dos dados que apresentam semelhanças, como por exemplo, informações referentes ao município de Curitiba, bem como avaliar as suas diferenças ao realizar comparações temporais.

Por fim, os resultados desta monografia podem contribuir com a metodologia para a construção de ambientes similares ao aqui proposto, uma vez que a criação de um data warehouse não implica necessariamente no melhor método de construção, e os fatores aqui analisados, como a qualidade dos dados e requisitos da área de negócio, podem impactam profundamente na eficiência do modelo apresentado.

Os resultados obtidos com o desenvolvimento do protótipo mostraram que a integração de dados governamentais é viável, e que, por meio deste trabalho, é possível dar continuidade aos usos das técnicas propostas através de trabalhos futuros, a fim de expandir a integração de dados entre diversos órgãos públicos, tornando as informações ainda mais relevantes e fundamentadas.

Seguindo no contexto de trabalhos futuros, a otimização das colunas para os tipos específicos do TCE-PR, pode trazer um excelente resultado, principalmente para agregações temporais e fatores numerais por exemplo. A mineração de dados também

pode ser explorada no escopo de dados deste trabalho, gerando uma contribuição adicional aos relatórios específicos construídos.

7. Referências

Araújo, L., Souza, J., Aumentando a Transparência do Governo por Meio da Transformação de Dados Governamentais Abertos em Dados Ligados. Revista Eletrônica de Sistemas de Informação, v. 10, n. 1, artigo 71, 2011.

Ballard, C., Data Mart Consolidation: Getting Control of Your Enterprise Information. IBM redbook. Rochester , 2005.

Ballard C., Herreman D., Schau D., Bell R., Kim E., Valencic A., Data modeling techniques for data warehousing. IBMRedBooks , 1998.

Barquim, R.; Edelstein, H., Planning and Design the Data Warehouse. Simon & Schuster, 1996.

Bastos, H., SIGA Brasil: Tecnologia da Informação a Serviço da Eficiência, Transparência e Controle Social do Gasto Público. Em: Senatus, v.7, n.1, p. 87-91, 2009.

Braga, L., Introdução à Mineração de Dados. E-Papers Serviços Editoriais, 2005.

Chaudhuri, S., Dayal, U., An overview of data warehousing and OLAP technology. Em: ACM Sigmod record, v. 26, n. 1, p. 65-74, 1997.

Ciferri, .C, Distribuição dos Dados em Ambientes de Data Warehousing: O Sistema WebD2W e Algoritmos Voltados à Fragmentação Horizontal dos Dados. Tese de Doutorado. Universidade Federal de Pernambuco. Centro de Informática, 2002.

Conhecendo o Tribunal / Tribunal de Contas da União. – 5. ed. Brasília : TCU, Secretaria-Geral da Presidência, 2011.

Dornelles, M., Iescheck, A., Análise da aplicabilidade da Infraestrutura Nacional de Dados Espaciais (INDE) para dados vetoriais em escalas grandes. Em: Bol. Ciênc. Geod., v. 19, n. 4, p. 667-686, Curitiba, 2013 .

Diniz, V., Como conseguir dados governamentais abertos. Em: III Congresso Consad de Gestão Pública. Brasília, 2009.

Faria, J., Artefatos da Semiótica Organizacional na Elicitação de Requisitos para Soluções de Data Warehouse. Tese de Mestrado. Universidade Estadual de Campinas. Instituto de Computação, 2006.

Fayyad, M.; Piatesky, G.; Smyth, P., From Data Mining to Knowledge Discovery: An Overview. Em: Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

Freitas, H., A informação como ferramenta gerencial: um telessistema de informação em marketing para apoio à decisão. Porto Alegre: Ortiz, 1993.

Golfarelli, M., Rizzi S., Data Warehouse Design – Modern Principles and Methodologies. McGraw-Hill, 2009.

Harper, F., Data warehousing and the organization of governmental databases. Em: Digital government: principles and best practices, p. 236, 2004.

Hu, X., Data Warehouse Technology and Application in Data Centre Design for E-government. INTECH Open Access Publisher, 2010.

Inmon, W., Building the Data Warehouse. John Wiley, 1996.

Irtishad A., Salman A., Pranas L., Development of a decision support system using data warehousing to assist builders/developers in site selection. Em: Automation in Construction, v. 13, n. 4, p. 525-542, 2004.

Imhoff, C., Galemmo N., Geiger J., Mastering Data Warehouse Design: Relational and Dimensional Techniques. Wiley, 2003.

Johnson, E., Jones, J., A Developer's Guide to Data Modeling for SQL Server: Covering SQL Server 2005 and 2008. Addison-Wesley Professional, 2008.

Kimball, R., The Data Warehouse Toolkit: Practical Techniques For Building Dimensional Data Warehouse. John Wiley & Sons, 1996.

List B., Bruckner R., Machaczek K., Schiefer J., A comparison of data warehouse development methodologies case study of the process warehouse. Em: Proceedings of the 13th International Conference on Database and Expert Systems Applications, P. 203-215, 2002.

Mundy, J., Thornthwaite W., Kimbal R.,The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset. John Wiley & Sons, 2008.

Mussi, C., Murahovschi, D., Bettini, G., & Kratz, L., Data Warehouse: A experiência da ANVISA. Em: Anais do IX CBIS – Congresso Brasileiro de Informática em Saúde, 2004.

Nebert, D., Developing Spatial Data Infrastructures: The SDI Cookbook. 2004.

PARANÁ. Tribunal de Contas. Lei Orgânica (Lei Complementar n. 113 de 15/12/2005) e Regimento Interno (Resolução n. 1 de 24/01/2006): versão seca e atualizada até 19 fev. 2015. Curitiba, 2015. 310 p.

Pires, F., Ambiente para extração de informações epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde. Tese de Doutorado. Universidade de São Paulo. Faculdade de Medicina, 2011.

Prather, C., Lobach, F., Goodwin, K., Hales, W., Hage, L., Hammond, E., Medical data mining: knowledge discovery in a clinical data warehouse. Proceedings of the AMIA Annual Fall Symposium, 101–105, 1997.

Rainardi,V. Building a data warehouse: with examples in SQL Server. John Wiley & Sons, 2008.

Ralha, C., Silva, C. A multi-agent data mining system for cartel detection in Brazilian government procurement. Em: Expert Systems with Applications, v. 39, n. 14, p. 11642-11656, 2012.

Rudra A., Yeo E., Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia. Em: Systems Sciences, Proceedings of the 32nd Annual Hawaii International Conference on. IEEE, 1999.

Santos, S., Almeida, L., Tachinardi, U., Gutierrez, A. Data warehouse para a saúde pública: estudo de caso SES-SP. Em: Anais do X Congresso Brasileiro de Informática em Saúde, p. 53-58, 2006.

Santos, W., Sistema de informação de custos do Governo Federal: modelo conceitual, solução tecnológica e gestão do sistema. IV Congresso de Gestão Pública, Brasília, 2011.

Silva, C., Ralha, C. Detection of cartel formation in government biddings using data mining agents. Revista Eletrônica de Sistemas de Informação, v. 10, n. 1, p. 1–8, 2011.

Sen, A., Sinha, A. A Comparison of Data Warehousing Methodologies: Using a common set of attributes to determine which methodology to use in a particular data warehousing project. Em: Communications of the ACM, v. 48, n. 3, p. 79-84, 2005.

United Nations. United Nations e-government survey 2010. Nova Iorque: UN Publishing Section, 2010. Disponível em: http://unpan1.un.org/intradoc/groups/public/documents/un/unpan038851.pdf. Acesso em: Junho de 2015.

Wu, M.-C., Buchmann, A.P. Research Issues in Data Warehousing. Em: Proceedings of The German Database Conference, pages 61-82, Germany, 1997.

APENDICE A – Scripts de Limpeza

```
#!/bin/bash
#Configuracao Inicial
ArquivoEntrada="dados_Processos.csv"
ArquivoSaida="insert_trâmiteProcessos.sql"
TabelaBD="TCC.trâmite_processos"
cat $ArquivoEntrada | sed 's/CIA, PEN/CIA PEN/g' | awk -F "," {'print
"("$1FS$2FS$3FS$4FS$5FS$6FS$7FS$8FS$9FS$10FS$11FS$12FS$13FS$14FS$
15FS$16FS$17FS$18")'"} | sed 's/  //g' | sed 's/"//g' | sed 's/,/","/g' | sed 's/(("/g' | sed
's/)/");/g' | awk -F ";" {'print "insert into $TabelaBD
(nmMunicipio,Entidade,TipoEntidade,nrProcesso,dsAssunto,dsAssuntoAgrupado,Loc
alizacao,LocalizacaoAgrupada,AnoAutuacao,AtoDecisao,Relator,Encerrado,cdMunici
pio,idAssunto,idRelator,idEntidade, dsEncaminhamento, DataRef) values "$1FS'} >
$ArquivoSaida

#!/bin/bash
#Configuracao Inicial
ArquivoEntrada="dados_dex.csv"
ArquivoSaida="insert_dex.sql"
TabelaBD="TCC.dex_determinacoes"
cat $ArquivoEntrada | sed 's/ART. 17,/17/g' | sed 's/NICO,/CO/g' | sed 's/,,/,null,/g' |
sed 's/)"/)/g' | sed 's/"//g' | sed 's/CARGO,/CARGO/g' | sed 's/,,/,null,/g' | awk -F ","
{'print
"(\""$1","$2","$3","$4","$5","$6","$7","$8","$9","$10","$11","$12","$13","$14","$15","$16
","$17"\");"'} | sed 's/,/","/g' | awk -F ";" {'print "insert into $TabelaBD
(Entidade,Gestor,TipoDeterminacao,idTipoDeterminacao,Determinacao,idDeterminac
ao,Prazo,DecisaoAto,DecisaoUnidade,DecisaoDataRegistro,idEntidade,idGestor,Baix
aAto,BaixaMotivo,BaixaUnidade,BaixaDataRegistro,nmMunicipio,cdMunicipio) values
"$1FS'} | sed 's/"null"/null/g' >  $ArquivoSaida

#!/bin/bash
#Configuracao Inicial
ArquivoEntrada="dados_municipios.csv"
ArquivoSaida="insert_municipios.sql"
TabelaBD="TCC.municipios"
cat $ArquivoEntrada | awk -F "," {'print
"("$1FS$2FS$3FS$4FS$5FS$6FS$7FS$8FS$9FS$10")'"} | sed "s/,/\",\"/g" | sed
"s/((\"/g" | sed "s/)/\")/g" | awk -F ";" {'print "insert into $TabelaBD
```

```
(cdUF,dsUF,cdMesoRegiao,dsMesoRegiao,cdMicroRegiao,dsMicroRegiao,cdMunicipi
o,dsMunicipio,cdIBGE,sgUF) values "$FS'} > $ArquivoSaida
#!/bin/bash
ANO=2015
#Configuracao Inicial
ArquivoEntrada=" consultaIpardes_anual.csv "
ArquivoSaida="insert_update_ipardes.sql"
TabelaBD="TCC.ipardes"
cat $ArquivoEntrada | sed 's/""/null/g' | sed 's/"-"/null/g' | awk -F ";" '{ \
if ($1=="Abatiá") {print "insert into $TabelaBD
(localidade,variavel,ano,valor,cdMunicipio) values ("$1","$2",$ANO,"$3",103"} \

\*

Bloco fixo de validação de munícipios.

*/

| sed 's/"\.\.\."/null/g' | sed 's/\.//g' | sed 's/TCCipardes/TCC.ipardes/g' > $ArquivoSaida
```

APENDICE B – Scripts Staging Área

```
CREATE TABLE [TCC].[ipardes](
	[localidade] [varchar](50) NOT NULL,
	[variavel] [varchar](100) NOT NULL,
	[ano] [int] NOT NULL,
	[valor] [float] NULL,
	[cdMunicipio] [varchar](5) NOT NULL,
	PRIMARY KEY ( cdMunicipio, variavel, ano ) );

CREATE TABLE [TCC].[dex_determinacoes](
	[Entidade] [varchar](50) NOT NULL,
	[Gestor] [varchar](50) NULL,
	[TipoDeterminacao] [varchar](50) NOT NULL,
	[idTipoDeterminacao] [varchar](50) NULL,
	[Determinacao] [varchar](1500) NULL,
	[idDeterminacao] [varchar](4) NULL,
	[Prazo] [varchar](4) NULL,
	[DecisaoAto] [varchar](50) NULL,
	[DecisaoUnidade] [varchar](4) NULL,
	[DecisaoDataRegistro] [datetime] NOT NULL,
	[idEntidade] [varchar](10) NULL,
	[idGestor] [varchar](10) NULL,
	[BaixaAto] [varchar](50) NULL,
	[BaixaMotivo] [varchar](50) NULL,
	[BaixaUnidade] [varchar](50) NULL,
	[BaixaDataRegistro] [datetime] NULL,
	[nmMunicipio] [varchar](50) NULL,
	[cdMunicipio] [char](5) NOT NULL
	PRIMARY KEY ( cdMunicipio, Entidade, TipoDeterminacao,
DecisaoDataRegistro ) );
```

```
CREATE TABLE [TCC].[municipios](
	[cdUF] [varchar](4) NULL,
	[dsUF] [varchar](20) NULL,
	[cdMesoRegiao] [varchar](10) NULL,
	[dsMesoRegiao] [varchar](50) NULL,
	[cdMicroRegiao] [varchar](10) NULL,
	[dsMicroRegiao] [varchar](50) NULL,
	[cdMunicipio] [char](5) NOT NULL,
	[dsMunicipio] [varchar](50) NULL,
	[cdIBGE] [varchar](10) NULL,
	[sgUF] [varchar](4) NULL
	PRIMARY KEY ( cdMunicipio ));

CREATE TABLE [TCC].[trâmite_processos](
	[nmMunicipio] [varchar](50) NULL,
	[Entidade] [varchar](100) NOT NULL,
	[TipoEntidade] [varchar](2) NULL,
	[nrProcesso] [varchar](100) NOT NULL,
	[dsAssunto] [varchar](50) NULL,
	[dsAssuntoAgrupado] [varchar](50) NULL,
	[Localizacao] [varchar](50) NOT NULL,
	[LocalizacaoAgrupada] [varchar](50) NULL,
	[AnoAutuacao] [varchar](20) NULL,
	[AtoDecisao] [varchar](10) NULL,
	[Relator] [varchar](100) NULL,
	[Encerrado] [varchar](20) NULL,
	[cdMunicipio] [varchar](10) NULL,
	[idAssunto] [varchar](10) NULL,
	[idRelator] [varchar](10) NULL,
	[idEntidade] [varchar](10) NULL,
	[dsEncaminhamento] [varchar](20) NULL,
	[DataRef] [varchar](20) NOT NULL,
PRIMARY KEY ( cdMunicipio, entidade, nrProcesso, DataRef ) );
```

APENDICE C – Scripts DW

```
CREATE DATABASE DW;

CREATE SCHEMA DW_TCE;

ON PRIMARY
(NAME = DW_Primary,
FILENAME = "\\psf\Home\Documents\TCC\DW_Prm.mdf",
 SIZE = 4MB,
 MAXSIZE = 10MB,
 FILEGROWTH = 1MB),

FILEGROUP DW_FG

(NAME = "DW_FG1",
FILENAME = "\\psf\Home\Documents\TCC\DW_FG1.ndf",
SIZE = 1MB,
MAXSIZE = 10MB,
FILEGROWTH = 1MB),

(NAME = "DW_FG2",
FILENAME = "\\psf\Home\Documents\TCC\DW_FG2.ndf",
SIZE = 1MB,
MAXSIZE = 10MB,
FILEGROWTH = 1MB)

LOG ON

(NAME = "DW_LOG",
FILENAME = "\\psf\Home\Documents\TCC\DW_LOG.ldf",
SIZE = 1MB,
MAXSIZE = 10MB,
FILEGROWTH = 1MB);

GO
```

```
CREATE TABLE [DW_TCE].[trâmite_processos](
	[cdMunicipio] [varchar](10) NOT NULL,
	[Entidade] [varchar](50) NOT NULL,
	[nrProcesso] [varchar](50) NOT NULL,
	[dsAssuntoAgrupado] [varchar](50) NULL,
	[dsAssunto] [varchar](50) NULL,
	[LocalizacaoAgrupada] [varchar](50) NULL,
	[Localizacao] [varchar](50) NOT NULL,
	[TipoEntidade] [varchar](2) NULL,
	[AnoAutuacao] [varchar](20) NULL,
	[idRelator] [varchar](10) NULL,
	[DataRef] [varchar](20) NOT NULL
	PRIMARY KEY ( cdMunicipio, entidade, nrProcesso, DataRef ) )
	ON "DW_FG";

GO


CREATE TABLE [DW_TCE].[dex_determinacoes](
	[cdMunicipio] [char](5) NOT NULL,
	[Entidade] [varchar](50) NOT NULL,
	[TipoDeterminacao] [varchar](50) NOT NULL,
	[Prazo] [varchar](4) NULL,
	[DecisaoDataRegistro] [datetime] NOT NULL,
	[DecisaoUnidade] [varchar](4) NULL,
	[BaixaMotivo] [varchar](200) NULL,
	[BaixaUnidade] [varchar](50) NULL,
	[BaixaDataRegistro] [datetime] NULL,
	[idGestor] [varchar](10) NULL,
	PRIMARY KEY ( cdMunicipio, Entidade, TipoDeterminacao,
DecisaoDataRegistro ) )
	ON "DW_FG";

GO
```

```
CREATE TABLE [DW_TCE].[ipardes](
	[localidade] [varchar](50) NOT NULL,
	[variavel] [varchar](100) NOT NULL,
	[ano] [int] NOT NULL,
	[valor] [float] NULL,
	[cdMunicipio] [varchar](5) NOT NULL,
	PRIMARY KEY ( cdMunicipio, variavel, ano) )
	ON "DW_FG";

GO

CREATE TABLE [DW_TCE].[municipios](
	[cdUF] [varchar](4) NULL,
	[dsUF] [varchar](20) NULL,
	[cdMesoRegiao] [varchar](10) NULL,
	[dsMesoRegiao] [varchar](50) NULL,
	[cdMicroRegiao] [varchar](10) NULL,
	[dsMicroRegiao] [varchar](50) NULL,
	[cdMunicipio] [char](5) NOT NULL,
	[dsMunicipio] [varchar](50) NULL,
	[cdIBGE] [varchar](10) NULL,
	[sgUF] [varchar](4) NULL
	PRIMARY KEY ( cdMunicipio ));
```