TRANSPORTES
www.revistatransportes.org.br
anpet
associação nacional de pesquisa e ensino em transportes
www.anpet.org.br

Desenvolvimento e aplicação de metodologia para consolidação de dados de pesquisas de tráfego no Departamento Nacional de Infraestrutura de Transportes

Maximiliano Martins de Faria[1], Julio César Barbieri[2], Geraldo Bonorino Xexéo[3], Glaydston Mattos Ribeiro[4], Nilo Flávio Rosa Campos Júnior[5], Leonel Antonio da Rocha Teixeira Júnior[6], Leonardo Roberto Perim[7]

[1]Universidade Federal do Rio de Janeiro, Brasil, maxfaria@cos.ufrj.br
[2]Universidade Federal do Rio de Janeiro, Brasil, juliobga@gmail.com
[3]Universidade Federal do Rio de Janeiro, Brasil, xexeo@cos.ufrj.br
[4]Universidade Federal do Rio de Janeiro, Brasil, glaydston@pet.coppe.ufrj.br
[5]Departamento Nacional de Infraestrutura de Transportes, Brasil, nilo.junior@dnit.gov.br
[6]Departamento Nacional de Infraestrutura de Transportes, Brasil, leonel.teixeira@dnit.gov.br
[7]Departamento Nacional de Infraestrutura de Transportes, Brasil, leonardo.perim@dnit.gov.br

Recebido:
15 de março de 2018
Aceito para publicação:
06 de junho de 2018
Publicado:
4 de novembro de 2018
Editor de área:
Bruno Vieira Bertoncini

Palavras-chaves:
Pesquisa de tráfego;
Consolidação de dados de tráfego;
Bases de dados de pesquisas.

Keywords:
Traffic survey;
Data consolidation process;
Traffic survey database.

DOI:10.14295/transportes.v26i3.1626

RESUMO
Este artigo apresenta os resultados de uma pesquisa-ação que teve como finalidade o desenvolvimento de um processo de consolidação de dados para pesquisas de tráfego no Departamento Nacional de Infraestrutura de Transportes (DNIT). A importância das pesquisas de tráfego para o planejamento do sistema viário nacional obrigou o DNIT a desenvolver um projeto que pudesse resgatar as informações de diversas pesquisas antigas de tráfego espalhadas por planilhas eletrônicas e outros tipos de documentos. O artigo apresenta a aplicação do processo de consolidação que resultou na criação de um banco de dados relacional possibilitando o resgate de dados de pesquisas antigas, distribuídas em planilhas eletrônicas que poderiam se perder ou se tornar inacessíveis com o passar dos anos.

ABSTRACT
This paper presents the results of an action research which purposes the development of a data consolidation process for traffic surveys in the National Department of Transportation Infrastructure (DNIT). The importance of traffic surveys for the national road system planning forced DNIT to develop a project that could retrieve the information from several old traffic surveys scattered through spreadsheets and other types of documents. This paper presents the application of this consolidation process and the creation of the relational database, which allows the retrieval of data from old traffic surveys.

1. INTRODUÇÃO

As pesquisas de tráfego são elementos importantes de um sistema viário de um país. Elas auxiliam no planejamento, análise e prospecção para novas estradas, expansões e melhoria das mesmas. No Brasil, o órgão responsável por todo o sistema viário de rodovias federais é o Departa-

mento Nacional de Infraestrutura de Transportes (DNIT). O DNIT é uma autarquia federal vinculada ao Ministério dos Transportes, Portos e Aviação Civil e tem como objetivo implementar a política de infraestrutura do Sistema Federal de Viação, compreendendo sua operação, manutenção, restauração, reposição, adequação de capacidade e ampliação mediante a construção de novas vias e terminais (DNIT, 2017).

Os marcos regulatórios do sistema viário brasileiro datam a partir da década de 70 (Martinovic *et al.*, 2016) e as pesquisas de tráfego fornecem informações valiosas sobre os aspectos operacionais do tráfego nos eixos viários, colaborando para as definições das políticas de transporte do País (DNIT, 2006). Desde dessa década, existem planos responsáveis pela contagem de tráfego nas rodovias nacionais, um destes planos é o "Plano Nacional de Contagem de Tráfego (PNCT)", que conta com postos de contagens permanentes e temporários espalhados pelo território nacional.

Após um longo período de inatividade, o PNCT foi retomado (a partir de 2013) com a contratação de serviços de contagem de tráfego em vários pontos da malha rodoviária federal (Júnior *et al.*, 2016). Muitos desses serviços foram originados a partir de diferentes contratos, onde não se observou uma preocupação com a padronização da contagem nem com a apresentação dos resultados e foram entregues em mídia eletrônica no formato de planilhas, documentos de textos e até mesmo por meio de figuras (fotos).

Com a retomada do novo PNCT e, consequentemente, o aumento das pesquisas de tráfego, se fez necessário padronizar as contagens para que as mesmas pudessem ser consolidadas em bases de dados únicas para assim, facilitar consultas através dos sistemas de informação do órgão. Com isso, um dos grandes desafios foi o de agregar as diversas fontes de contagens de tráfego existentes com as fontes atuais já consolidadas. A ideia do órgão foi criar uma base centralizada, que consolidasse as pesquisas antigas com as pesquisas atuais a partir de 2013. Este trabalho apresenta a aplicação de um processo desenvolvido por pesquisadores da Coppe/UFRJ, através de uma parceria com o DNIT que automatiza, analisa e consolida todas as informações de diversas pesquisas antigas de tráfego do órgão espalhadas por planilhas eletrônicas e outros tipos de documentos criando, a partir destas, bases de dados normalizadas e centralizadas.

Este artigo será apresentado em cinco seções, sendo esta a primeira. Na Seção 2 são discutidos trabalhos relacionados ao problema em questão. A Seção 3 descreve a metodologia utilizada que tem como resultado a obtenção de bases de dados centralizadas. Em seguida, a Seção 4 mostra o desenvolvimento e a aplicação da metodologia aos dados reais do DNIT e, por fim, a Seção 5 que apresenta as conclusões finais.

2. TRABALHOS RELACIONADOS E CONCEITOS BÁSICOS

Este trabalho propõe um processo auditável que analise, transforme e consolide os dados de pesquisas antigas de tráfego em bancos de dados relacionais (Codd, 1970). Foi feita uma ampla revisão sobre o tema na busca por tecnologias que pudessem auxiliar a construção de um processo de consolidação de dados. Nesta pesquisa foram encontrados muitos trabalhos, relativos a tratamento de dados de tráfego, utilizando as tecnologias de *data warehouse* (DW) e processos de ETL (*Extract – Transform – Load*).

Data warehouse (DW) é um banco de dados integrado composto por diversas fontes de dados, muito usado para consolidar informações de diversas bases e armazená-las de forma centralizada, o que facilita as consultas por sistemas de uma organização e melhora as tomadas de

decisões (Kimball and Ross, 2013; Moalla *et al.*, 2017; Jarke *et al.*, 2013; Mukherjee and Kar, 2017).

Existem diversas abordagens, tipos, modelos e maneiras de se montar um DW, mas uma discussão detalhada sobre não é o foco deste trabalho. Pode-se destacar algumas características principais como: a orientação do sistema DW, sua arquitetura (ou esquema), dimensão dos dados, entre outras.

Baseado no objetivo deste trabalho, optou-se por construir um ambiente DW de orientação não volátil, com o esquema *snowflake* e multidimensional (Mukherjee and Kar, 2017). Na orientação não volátil os dados oriundos da consolidação uma vez carregados no banco de dados não poderão ser apagados. Quanto ao esquema *snowflake* trata-se de como as tabelas serão criadas (Moalla *et al.*, 2017). Neste esquema existe uma tabela de fatos com chaves identificadoras das tabelas dimensões. Essas chaves são a base dos relacionamentos entre essas tabelas (Figura 1).

Uma das técnicas mais importantes dentro de um ambiento DW é o ETL. O ETL é processo composto de técnicas para extração de dados de diversas fontes, limpeza e transformação segundo um padrão desejado, para então serem carregados em banco de dados (Gill and Singh, 2014; Prema and Pethalakshmi, 2013). Este processo será responsável por toda a manipulação nos dados das pesquisas legadas e por último pela inserção destas pesquisas no banco de dados.

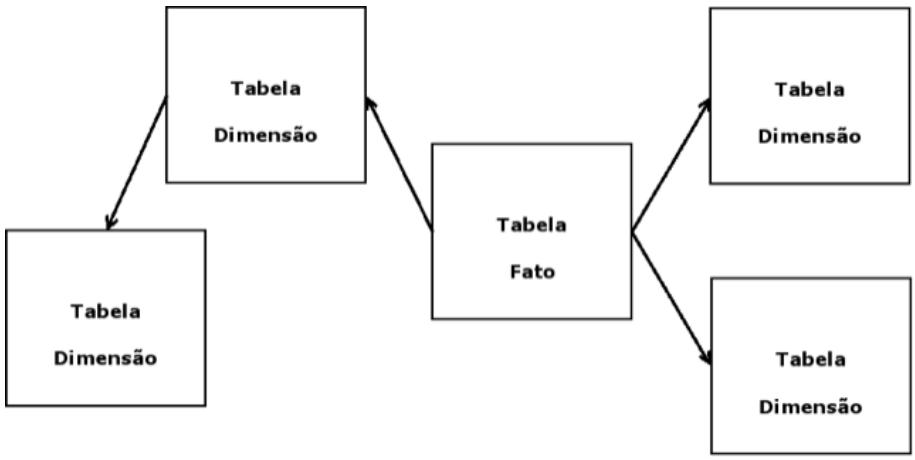

Figura 1. Relacionamento entre chaves da tabela fato com diversas tabelas dimensão do banco a ser consolidado. Esquema *snowflake* em um ambiente DW

As fontes de dados podem ter muitos formatos como: arquivos textos, arquivos XML, tabelas relacionais, arquivos de *logs*, planilhas eletrônicas entre outros (Prema and Pethalakshmi, 2013). A diversidade de fontes de dados é uma das caraterísticas deste trabalho e as duas tecnologias encontradas, *data warehouse* e ETL, atendem perfeitamente seus objetivos, tanto no quesito técnico como no embasamento teórico.

A opção por usar DW e as técnicas ETL para a construção de uma base de dados consolidada se mostrou acertada, uma vez que existe ampla literatura sobre o tema e muitos relatos de sucesso na indústria, mas importante ressaltar que como todo sistema computacional existem vantagens e desvantagens. A principal vantagem e ponto decisivo é poder criar um processo claro para a consolidação de dados legados e condenados ao esquecimento, como desvantagem pode-se citar a demora da conclusão do processo como um todo, tanto na análise das fontes de

dados (planilhas e documentos de pesquisas de tráfego legadas) quanto na execução das técnicas ETL para a carga no banco. Na Figura 2 é apresentado uma visão geral de um ambiente DW, fonte de partida para este trabalho.

A definição da arquitetura do ambiente a ser usada foi o ponto de partida para este trabalho, como passo seguinte foi execução de uma ampla pesquisa na literatura acadêmica que apresentasse trabalhos e problemas semelhantes aos descritos aqui. A seguir são apresentados estudos técnicos científicos com a aplicação de *data warehouse* e ETL em pesquisas de tráfego.

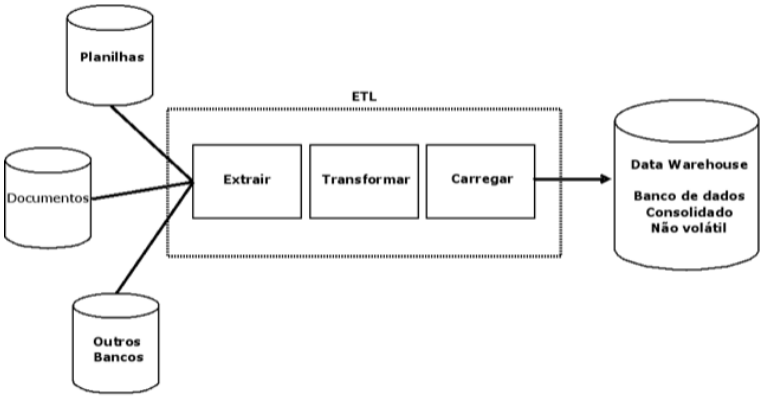

Figure 2. Visão geral de um ambiente de Data Warehouse

Em Bitar (2016) os pesquisadores, através de atividades de ETL, consolidaram informações de bases de dados de tráfego do estado de Oklahoma, Estados Unidos, para criação de um *data warehouse.* A partir deste modelo os pesquisadores executaram análises estatísticas, filtragem e correção de dados para obtenção de informações sobre padrões de tráfego das estradas que apoiassem as decisões das agências de transportes quanto à melhoria do tráfego na cidade.

O trabalho de Tao *et al.* (2015) propõe um sistema de gerenciamento de dados que melhore as capacidades de armazenamento, desempenho, acesso e integração aos dados de tráfego. Os pesquisadores, do Laboratório de Operação e Segurança de Tráfego da Universidade Wisconsin em Madison (EUA), propuseram uma reformulação em sua base de dados estatal de tráfego, através de técnicas de ETL, para a criação de um novo *data warehouse* que pudesse aprimorar os armazenamentos de dados, diminuindo os intervalos de amostras de tráfego de 5 para 1 minuto.

Em Zhao *et al*. (2011) foi proposto um modelo para criação de um Sistema de Transporte Inteligente (ITS) que centralizasse informações de tráfego coletados a partir de dispositivos coletores de viagens. Para isso o trabalho propõe a criação de um *data warehouse* ITS que centralize os dados de tráfego e possa ser usado em consulta e no apoio a tomada de decisões.

Sun (2011) realizou um estudo pelo Departamento de Engenharia Civil e Ambiental da Universidade de Melbourne para melhoria da manutenção do pavimento rodoviário. O pesquisador descobriu que os dados nacionais sobre gastos do setor rodoviário poderiam ser melhorados com uma abordagem de gerenciamento do conhecimento através da captura de dados sobre tráfego e sua consolidação e o uso em sistemas de *data warehouse*.

Por fim, Von Brown *et al*. (2011) executaram uma ampla pesquisa em diversos estados Norte Americanos sobre como eram gerenciados seus dados sobre segurança de tráfego. Alguns estados responderam a pesquisa indicando o uso de sistemas *data warehouse* para centralização e análise de dados como, por exemplo, Virginia, Maine e Michigan.

Nota-se o vasto uso de tecnologias *data warehouse* e processos ETL na consolidação, análise, limpeza e transformação de dados. Na maioria dos artigos a intenção é a de apoiar à tomada de decisão e de dar transparência no trato dos dados. Mas nenhum dos artigos cita o resgate de pesquisas legadas distribuídas em fontes como arquivos textos e planilhas eletrônicas. Neste sentido isto torna este trabalho de certa forma inovador e com uma rica contribuição.

3. METODOLOGIA

Para o desenvolvimento deste trabalho foi necessário a criação de um processo que pudesse agregar as diversas fontes de dados antigos (planilhas eletrônicas, entre outros) e em seguida serem analisadas e consolidadas em bases de dados relacionais. Para alcançar este objetivo, a metodologia foi dividida em duas partes: Na primeira foi executado um trabalho de pesquisa-ação que pudesse auxiliar na busca por soluções e a segunda foi dedicada ao desenho do processo de consolidação com etapas claras e que pudessem ser auditadas.

A metodologia usada neste trabalho, para busca das soluções, foi baseada na pesquisa-ação que pode ser descrita como uma pesquisa que executa uma ação e ao mesmo tempo uma investigação (Reason e Bradbury, 2001). A pesquisa-ação, dentro de uma pesquisa, é uma forma de investigação interna cujo objetivo é resolver problemas, aperfeiçoar práticas e propor soluções (Espadas *et al.*, 2008). A mesma está baseada em um processo de inteligência crítica para dar forma a uma ação, desenvolvendo capacidades para reestruturações e propondo novas soluções que sejam necessárias (Herbert Altrichter *et al.*, 2002).

A pesquisa-ação deste trabalho foi elaborada com base em processos de tratamento e consolidação de base de dados encontrados na literatura. O resultado da revisão da literatura e dos trabalhos relacionados (Seção 2) foi a criação de um processo, baseado em tecnologia de *data warehouse* e atividades de ETL de dados. A Figura 3 apresenta uma visão da metodologia da pesquisa-ação utilizada no desenvolvimento do processo deste trabalho.

Após a finalização da pesquisa-ação, desenvolve-se o desenho do processo aqui denominado de "**Processo de Consolidação**". O objetivo é que através deste processo pudessem ser extraídos os dados das pesquisas antigas e assim criar um banco de dados relacional relativo a cada pesquisa de tráfego antiga.

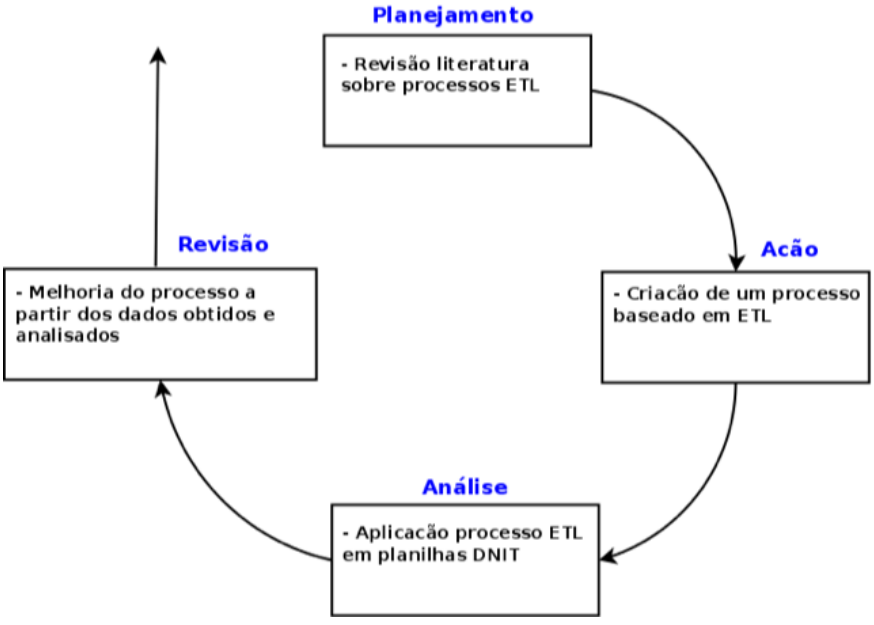

Figura 3: Fases da pesquisa-ação utilizada na criação do processo utilizado na consolidação dos dados

O processo planejado foi baseado em conceitos de *data warehouse* e nas atividades de extração, transformação e carga de dados (ETL) conhecido de ambientes de *data warehouse* (Dhakal, 2014; Inmon e Hackathorn, 1994). O processo de consolidação foi dividido em cinco etapas, são elas: **Modelagem de dados, Criação do banco de dados, Limpeza de dados, Atividades de ETL e Verificação de dados**. Na Figura 4 é apresentada uma visão geral do processo construído e em seguida a descrição de cada etapa.

3.1 Modelagem da base de dados

Durante essa etapa, cada arquivo recebido dos projetos é analisado minuciosamente. O objetivo é a criação de um modelo de dados para cada um dos projetos. Os modelos de dados resultantes são baseados naqueles já estabelecidos pelo DNIT, permitindo assim, a compatibilidade entre os modelos de cada projeto.

O objetivo deste trabalho é resgatar pesquisas de tráfego antigas condenadas ao esquecimento, muito destas oriundas de editais que não se preocupavam com os dados e sim apenas em fazer a pesquisa de tráfego. Isto criou uma enorme falta de padronização nas pesquisas o que dificultava ainda mais as análises e possíveis consolidações.

Neste sentido esta etapa vem cumprir o papel de analisar cada projeto, seus dados e como estes poderão ser armazenados em banco de dados. Tomou-se como base as pesquisas atuais de tráfego desenvolvidas pelo DNIT e seus modelos de dados. A partir destes pode-se implementar um modelo de dados padrão para os projetos, vale ressaltar que o mesmo não era estático, pelo contrário, na medida que as etapas do processo de consolidação (Figura 4) avançavam, mudanças no modelo podiam acontecer e consequentemente refletindo nas tabelas do banco de dados do projeto.

Na Figura 5 é apresentado o modelo padrão usado nas pesquisas de tráfego volumétricas, aqui são apresentadas apenas as principais tabelas. O modelo segue a estrutura básica do esquema *snowflake* de um ambiente DW (Figura 1) conforme descrito na seção 2.

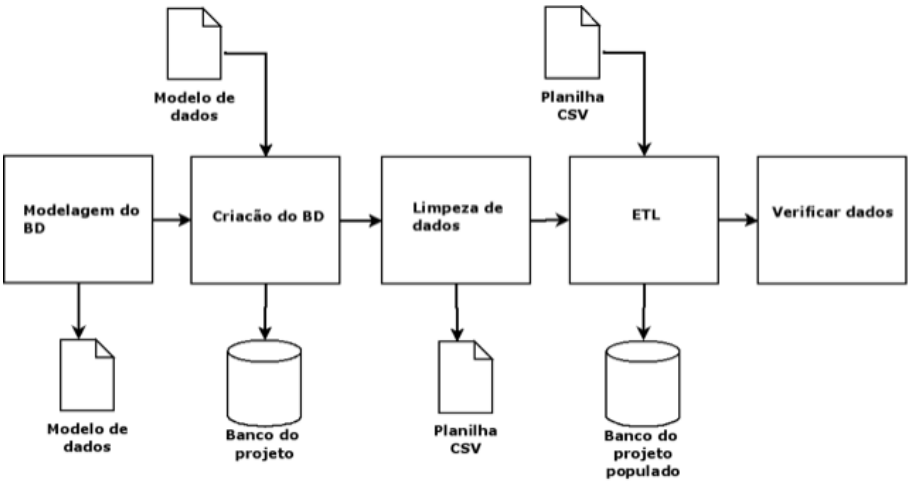

Figura 4. Processo de consolidação dos dados das pesquisas antigas

3.2 Criação do banco de dados

Após a modelagem, nesta etapa é criado o banco de dados a partir do modelo. Este banco receberá os dados do projeto no fim do processo. Durante o decorrer do processo, o modelo pode

ser alterado para refletir mudanças encontradas nas etapas seguintes. Cada projeto terá seu banco relacional único, mas como todos seguem um modelo padrão do DNIT, no fim do processo de consolidação, o órgão poderá juntar todos os bancos dos projetos em um único banco.

A criação do banco é baseada no que foi definido na modelagem de dados, após a modelagem é gerado arquivos de configurações com comandos em SQL (Date and Darwen, 1997) contendo todas as informações sobre as tabelas e seus campos. A partir destes arquivos SQL os bancos são criados no software gerenciador de banco de dados. O software utilizado para o banco de dados foi o MySQL versão 5.6 (MySQL, 2001), este é um software livre e atendeu todos os objetivos do projeto, muito por sua portabilidade, escalabilidade e facilidade no manuseio. Na Figura 6 é apresentado uma parte de uma arquivo SQL para a criação de uma tabela do banco.

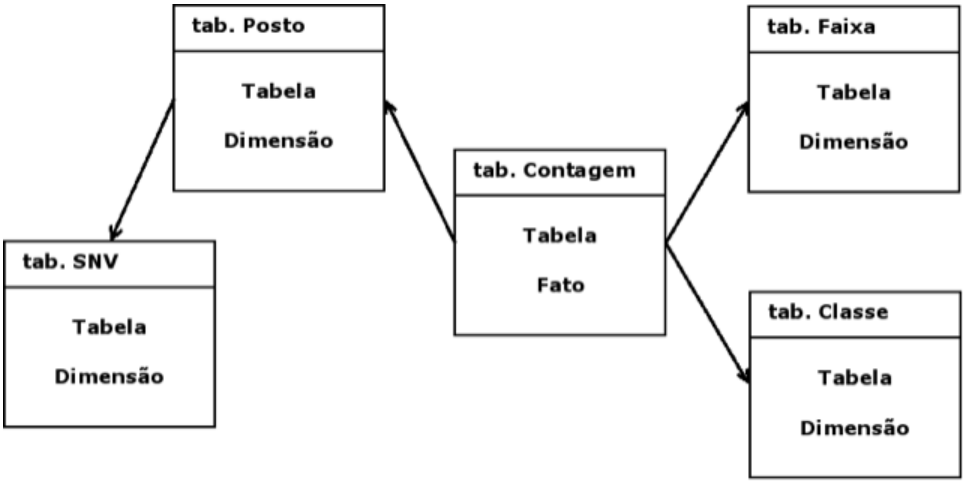

Figura 5. Modelo de dados (resumido) utilizado para pesquisas de tráfego volumétrica

```
-- -----------------------------------------------------
-- Table `projeto#1`.`TB_CONTAGEM_VOLUMETRICA`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `projeto#1`.`TB_CONTAGEM_VOLUMETRICA`
  `id_contagem_volumetrica` INT(11) NOT NULL AUTO_INCREMENT,
  `id_classe` INT(11) NULL DEFAULT '0',
  `id_posto` INT(11) NULL DEFAULT NULL,
  `id_faixa` INT(11) NULL DEFAULT NULL,
  `dt_hr_contagem` DATETIME NULL DEFAULT NULL,
  `quantidade` INT(11) NULL DEFAULT NULL,
  `segmento` VARCHAR(45) NULL DEFAULT NULL,
  `sentido` VARCHAR(45) NULL DEFAULT NULL,
  PRIMARY KEY (`id_contagem_volumetrica`),
```

Figura 6: Arquivo SQL para criação de uma tabela **CONTAGEM** no banco

3.3 Limpeza de dados

Nesta etapa as diversas planilhas eletrônicas de um projeto são convertidas em arquivos *comma-separated values* (CSV) com o mesmo formato. O formato destes arquivos está baseado nos modelos criados na etapa anterior e pode variar de projeto para projeto, entretanto as bases resultantes de cada projeto devem possuir estruturas de tabelas que sejam similares, para que possa existir compatibilidade entre as bases. A Figura 7 apresenta um exemplo de arquivo CSV resultante do processo inicial de limpeza de dados.

Esta etapa é uma das mais importantes, aqui pela primeira vez os dados das pesquisas de tráfego, sejam em planilhas ou outros documentos, são importados seguindo o padrão do modelo de dados. Também nesta fase, podem ocorrer mudanças no modelo, pois dependendo do projeto podem ter informações que não foram contempladas no modelo padrão (Figura 4) e só nesta fase foram descobertas.

Outro aspecto importante desta fase é a limpeza de dados com problemas, dados faltantes e dados ilegíveis. Muitas vezes as pesquisas continham dados incompletos, como por exemplo o UF da pesquisa. Outro problema também recorrente era a falta de padronização na classe dos veículos, muitas vezes a descrição da classe de veículo variava dentro da mesma pesquisa. Então foram necessários criar pequenos programas que automatizassem a correção destes erros, em outros casos quando a automatização não foi possível as correções tiveram que ser manuais dentro das próprias pesquisas. Ao fim desta fase, o arquivo tipo CSV (Figura 7) estava pronto para ser usado pelo processo ETL.

```
1  Date/Time   Class #1    Class #2    Class #3    Class #4    Class #5    Class #6    Class #7
   Class #8    Class #9    Class #10   Class #11   Class #12   Class #13   Sentido Nome_Sentido
   Latitude    Longitude   KM  UF  BR
2  28/7/2014 00:00 0   1   0   0   0   0   0   0   0   0   0   0   SENTIDO LESTE-OESTE CAETITÉ-BA
           109 BA  30
3  28/7/2014 01:00 0   1   0   0   1   0   0   0   0   0   0   0   SENTIDO LESTE-OESTE CAETITÉ-BA
           109 BA  30
4  28/7/2014 02:00 0   1   0   0   0   0   0   0   0   0   0   0   SENTIDO LESTE-OESTE CAETITÉ-BA
           109 BA  30
5  28/7/2014 03:00 0   3   0   0   0   1   0   0   0   0   0   0   SENTIDO LESTE-OESTE CAETITÉ-BA
           109 BA  30
```

Figura 7: Exemplo do arquivo CSV após a execução da limpeza de dados no mesmo.

3.4 Construção da transformação

Nesta etapa ocorrem as transformações dos dados em bases de dados. As transformações se baseiam em conceitos de ETL (Kakish e Kraft, 2012; Prema and Pethalakshmi, 2013; Titirisca, 2013) e recebem todos os arquivos CSV da etapa anterior como entrada. A base de dados resultante tem fundamento no modelo criado na etapa anterior de modelagem da base de dados. Nesta etapa, para as transformações, utilizou-se um *software* de apoio ao ETL. Neste trabalho o *software* escolhido foi o *Pentaho Data Integration* (Casters *et al.*, 2010).

3.4.1 Extract, Transform, Load (ETL)

Segundo Casters *et al*. (2010), Satkur (2013) e Kakish e Kraft (2012), sistemas de ETL são sistemas baseados em extração de dados de diversas fontes externas, transformações conforme as regras de negócio e carga, geralmente, em um *data warehouse*. Em geral, o processo de ETL é dividido em 3 atividades:

- ***Extract* (Extração)**: Etapa que compreende a conexão com as fontes de dados, sua extração e disponibilização para as demais etapas do processo. Esta atividade é executada dentro do *software Pentaho Data Integration,* aqui é criada uma conexão com o arquivo CSV da etapa anterior. Esta conexão abre o CSV e seleciona linha a linha importando-as para dentro do software.
- ***Transform* (Transformação)**: Etapa que envolve a aplicação de regras/funções nos dados extraídos com o objetivo de derivar os dados que serão carregados. Aqui cada linha importada na extração passa por uma série de regras que irão executar transformações

em seus dados. Essas transformações executam as mais diversas tarefas, desde de simples conferências em UF, conferências nos KM (quilômetro onde foi feito a pesquisa), conferências em SNV (Sistema Nacional de Viação), padronizações das classes de veículos entre outras transformações. Na Figura 8 segue um exemplo de transformação em um projeto, pode-se notar que as transformações são apresentadas como fluxos ramificados por onde cada linha do arquivo CSV seguirá.

- ***Load* (Carga)**: Etapa que compreende a carga dos dados transformados em um *Data Warehouse.*

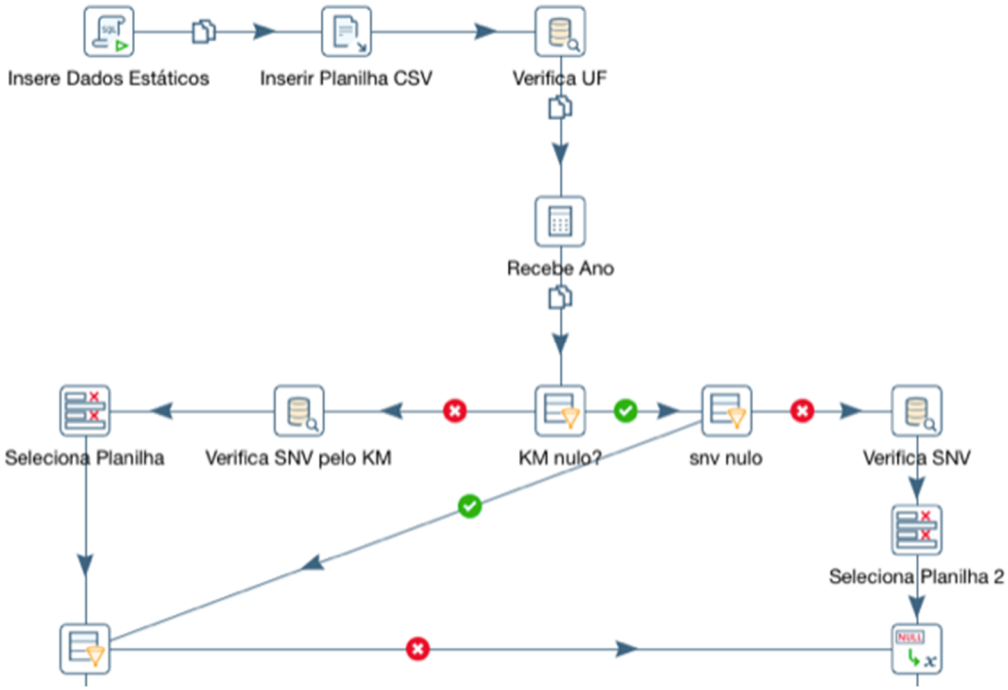

Figura 8. Exemplo de transformações na UF, SNV e KM nos dados de um projeto

A etapa de transformação permite a modificação do fluxo de dados de diferentes formas. As transformações mais comuns realizadas nos dados são: seleção e quebra de colunas, limpeza de dados, codificação de valores, derivações de novos valores e junções. A Figura 9 apresenta uma consulta em um banco de dados relacional após a execução de toda a etapa da transformação.

#	id_contagem_volumetrica	sg_sigla	sg_sentido	dt_hr_contagem	quantidade	codigo
1	146735	M	D	2014-02-13 00:00:00	0	153BTO0210
2	147737	P2	D	2014-02-13 00:00:00	0	153BTO0210
3	147236	P1	D	2014-02-13 00:00:00	0	153BTO0210
4	148238	P3	D	2014-02-13 00:00:00	0	153BTO0210
5	148739	O1	D	2014-02-13 00:00:00	2	153BTO0210
6	149741	O3	D	2014-02-13 00:00:00	1	153BTO0210
7	149240	O2	D	2014-02-13 00:00:00	0	153BTO0210
8	150242	C1	D	2014-02-13 00:00:00	0	153BTO0210
9	152246	C5	D	2014-02-13 00:00:00	0	153BTO0210
10	153248	R2	D	2014-02-13 00:00:00	0	153BTO0210

Figura 9: Exemplo de dados, a partir de arquivos CSV, já transformados em banco de dados relacional.

3.5 Verificação de dados

Esta é a última etapa do processo. Aqui os dados importados para a base de dados são comparados com as planilhas originais. O objetivo é encontrar possíveis inconsistências ou erros que não foram corrigidos durante o processo. Esta etapa é de grande importância, pois, a partir de seus resultados, alterações nas etapas anteriores podem ser necessárias. Esta etapa age como um mecanismo de validação de todo o processo.

Para este processo de verificação, foram utilizadas amostras aleatórias simples de uma população composta por pesquisas de tráfego. A amostra retirada foi comparada com os dados originais das planilhas CSV através de um pequeno programa. Foram utilizados conceitos estatísticos simples para se calcular a amostra, desde o conceito de amostra significativa, população, até erro amostral (neste trabalho foi utilizado erro de 5%) e nível de confiança (no trabalho foi utilizado um nível de confiança de 95%) (Omair, A. 2014).

4. DESENVOLVIMENTO E APLICAÇÃO DO PROCESSO

A Seção 3 descreveu a metodologia utilizada para a criação do processo de consolidação dos dados das pesquisas de tráfego antigas do DNIT. Nesta seção apresenta-se a aplicação deste processo em dois projetos de pesquisa de tráfego do órgão. O DNIT disponibilizou dados de oito projetos com mais de mil itens, entre planilhas eletrônicas, documentos textos, fotos e arquivos do tipo *Portable Document Format* (PDF). O processo de consolidação (Figura 4) foi aplicado principalmente nas planilhas eletrônicas, pois são nestas onde residem os dados de pesquisa. Mas é importante ressaltar que os outros arquivos (documentos textos, fotos entre outros) também foram analisados, já que, muitas vezes os dados que deviam estar contidos nas planilhas foram separados em outros tipos de arquivos.

Em todos os projetos foram encontrados pesquisas de tráfego de no mínimo dois tipos: Pesquisa de Contagem Volumétrica Classificatória e Pesquisa de Origem e Destino (OD). Para o desenvolvido deste artigo, foram separados dois projetos, ambos já finalizados, aqui denominados de Projeto #1 e Projeto #2. A Tabela 1 sintetiza os dados destes projetos.

Tabela 1: Dois projetos onde o processo de consolidação foi aplicado

Projeto	Num. De Postos	Data Pesquisa	Tipo pesquisa	Qtd. Pesq. Volumétrica	Qtd. Pesq. OD
Projeto #1	28	Entre 12/03/2014 à 13/11/2014	Volumétrica e OD	191.568	297
Projeto #2	73	Entre 13/02/2014 à 19/03/2016	Volumétrica e OD	2.399.808	19.826

4.1 Desenvolvimento do processo

A partir do planejamento do processo de consolidação (Seção 3), esta subseção demonstra como o processo foi desenvolvido e qual seu objetivo. A visão resumida do desenvolvimento do processo e sua aplicação está apresentada na Figura 10, importante ressaltar que todo este foi baseado na metodologia desenvolvida pela equipe do projeto. O desenvolvimento do processo foi dividida em duas grandes fases:

Fase adaptação dos dados:

- Este é o grupo das fases iniciais do processo. Aqui encontram-se as atividades de: **Modelagem, Criação da base de dados e Limpeza dos dados** (Seções 3.1, 3.2 e 3.3). Nele

acontece a modelagem do banco de dados a partir da análise de todas as planilhas do projeto, também é executado um processo de limpeza para eliminar problemas de caracteres irreconhecíveis, campos sem dados, *outliers*, datas erradas, entre outros; e

- fim desta fase resulta em um arquivo CSV com todos os dados de cada planilha, ou seja, geram-se arquivos CSV para cada planilha de dados.

Fase transformação dos dados:

- Este grupo compreende as atividades ETL (Seções 3.4 e 3.5) desde importação, transformação e criação da base de dados relacional, a partir dos arquivos CSV das etapas anteriores;
- Cada arquivo CSV sofre diversas transformações e verificações até se tornar um banco de dados relacional;
- Dentro das atividades ETL que acontecem nesta fase, várias ações sob os dados dos arquivos CSV são executadas. Podem-se destacar ações de verificações de consistência de datas, verificações de codificação do Sistema Nacional de Viação (SNV) (Casa Civil, 2011), preenchimento de dados faltantes, verificações de estados e municípios, entre outras;
- No final das atividades de ETL, um banco de dados final recebe os dados de todos os arquivos CSV da etapa anterior. Cada projeto terá um banco de dados com todos os dados de pesquisa oriundos das planilhas; e
- Ao fim desta fase, os dados importados para o banco de dados são confrontados com as planilhas originais. São feitas verificações por amostragem, tanto para pesquisas OD quanto para pesquisas Volumétricas. As amostras tem como erro amostral 5% e nível de confiança 95%. Caso inconsistências sejam encontradas, todo o processo deve ser revisado. De acordo com o resultado desta etapa, a modelagem dos dados da fase inicial pode sofrer mudanças.

Com o processo desenvolvido, o mesmo se encontra pronto para ser aplicado nos projetos de pesquisas de tráfego antigas. O processo é executado de forma interativa e por projeto, ou seja, cada projeto passa por todo o processo de consolidação. Entretanto, cada um pode sofrer mudanças únicas, o que acarretará em diferenças no nível de detalhes de projeto para projeto, mas na visão geral, todos eles seguiram igualmente todo o processo desenhado e descrito na Seção 3.

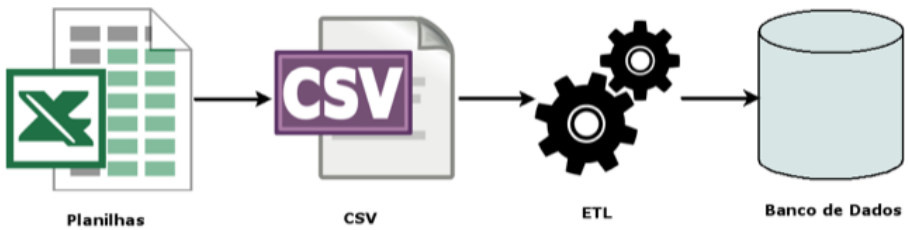

Figura 10. Visão resumida do Processo de Consolidação das planilhas recebidas por projeto

4.2 Aplicação do Processo de Consolidação

O processo desenvolvido foi aplicado nas planilhas dos dois projetos que envolveram pesquisas de tráfego, conforme detalhamento apresentado na Tabela 1.

4.2.1 Projeto #1

Os arquivos deste projeto são relativos as pesquisas que ocorreram durante o ano de 2014. Essa pesquisa fez parte de um projeto para implantação de um sistema de gerência de pavimentos (SGP) integrado às soluções de Geoprocessamento via Web para a realização das atividades referentes aos Estudos Preliminares da fase de Estudos e Projetos Rodoviários. Para isso foram realizados estudos preliminares que incluíram as pesquisas de campo pontuais contidas nos arquivos recebidos (planilhas, por exemplo).

Arquivos Recebidos

Para este projeto foram recebidos mais de 450 arquivos (planilhas, fotos e etc.), que possuíam mais de 191.000 veículos contados e classificados e 290 entrevistas de origem e destino.

Aplicação do Processo de Consolidação

A aplicação do processo de consolidação (Seções 3 e 4.1) neste projeto seguiu todas as etapas já descritas. Nas etapas iniciais aconteceram as análises de todas as planilhas do projeto, onde os dados foram adaptados e convertidos em arquivos CSV. Nas etapas finais, os arquivos CSV serviram de entrada para as transformações dos dados e para a criação do banco de dados relacional consolidado.

O banco de dados relacional resultante possui uma estrutura padrão, pré-estabelecida, que todos os projetos seguiram. Na Figura 11 é apresentado o modelo de dados do banco final do Projeto #1. Por questões de compatibilidade, o modelo criado segue as normas de modelos de dados definidos pelo DNIT. A intenção é a de que os bancos relacionais resultantes possam ser consolidados em um único modelo de dados.

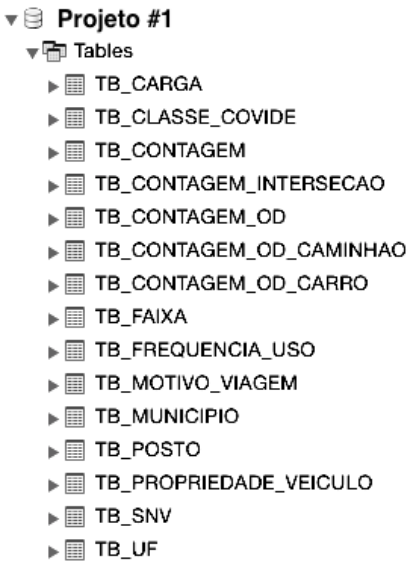

Figura 11: Modelo de dados final Projeto #1, resultante do Processo de Consolidação.

Durante o processo de consolidação alguns desafios foram encontrados. As classes de veículos utilizadas nas contagens volumétricas do projeto, por exemplo, estavam diferentes das classes adotadas como padrão pelo DNIT. Nestes casos, durante as transformações de dados, foram executados alinhamentos entre as classes encontradas com as classes padronizadas.

Outro problema detectado é de que o padrão de contagens volumétricas estabelecido pelo DNIT indica contagens de 15 em 15 minutos, porém as contagens encontradas no Projeto #1

foram realizadas de hora em hora. Dados faltantes também foram um desafio para a consolidação dos dados. Foi possível notar que muitas planilhas não possuíam códigos SNV, quilômetro ou coordenadas dos postos de pesquisas. Todos estes problemas foram tratados na etapa de transformações, por exemplo a questão dos intervalos de tempo foram mantidos os tempos usados nas pesquisas originais (planilha), ou seja, neste projeto o intervalo de tempo entre contagens de veículo passou a ser de hora em hora. A questão de dados faltantes, no caso ausência do SNV, foi possível deduzi-lo através de outros dados na planilha como por exemplo pela quilometragem do local da pesquisa. Já em casos que não tinham quilometragem do local na planilha, o processo teve que ser paralisado para que fossem feitas intervenções manuais como por exemplo abrir mapas eletrônicos e localizar o local da pesquisa para então se encontrar o SNV.

Ao fim do processo um banco de dados relacional foi gerado e importado para um Sistema de Gerenciamento de Banco de Dados (SGBD) (Merriam-Webster, 2018), no caso deste trabalho foi usado MySQL (Zawodny e Balling, 2008). A verificação da consistência dos dados importados foi executada por meio de um processo de amostragem, onde para cada tipo de pesquisa (Volumétrica e OD) foi calculada uma amostra representativa destas, com erro amostral de 5% e nível de confiança de 95%. Para o cálculo da amostra, este trabalho baseou-se no artigo de Omair, A. 2014. Na Tabela 2 são apresentados os valores das amostras utilizadas.

Tabela 2: Número de amostras utilizada por tipo de pesquisa no Projeto #1.

Tipo de pesquisa	Qtd. de registros	Tamanho de amostra
Volumétrica	2399808	385
Origem-Destino (OD)	297	168

Para extrair a amostra, foram feitas consultas simples no banco para retirar o número desejado de registros. Os registros extraídos foram comparados com os dados na planilha CSV através de um pequeno programa desenvolvido para este fim. No Projeto #1 todas as consultas executadas estavam condizentes com os dados encontrados nas planilhas originais.

4.2.2 Projeto #2

Os arquivos inclusos neste projeto são relativos a pesquisas de tráfego que ocorreram durante os anos de 2014 à 2016. As pesquisas fazem parte de um grande projeto de contratação de empresas especializadas para execução de estudos de planejamento da infraestrutura de transportes sob supervisão do DNIT.

Arquivos Recebidos

Foram recebidos mais de 230 arquivos (planilhas, fotos e etc.), onde estavam contidos mais de 2.300.000 de veículos contados e classificados e mais de 19.800 entrevistas de origem e destino.

Aplicação do Processo de Consolidação

Conforme executado no Projeto #1, o mesmo processo foi repetido no Projeto #2. Inicialmente foram analisadas todas as planilhas, com execução das adaptações e conversões para arquivos CSV. Em seguida, os arquivos gerados sofreram as transformações e foi criado o banco de dados relacional do projeto.

Como no Projeto #1, o modelo de dados gerado para o Projeto #2 também seguiu o modelo de dados pré-estabelecido pelo DNIT, já que a compatibilidade entre os projetos foi uma questão importante para a consolidação dos dados. Entretanto, algumas adaptações foram realizadas

em virtude da natureza distinta de determinados dados dos projetos.

Na Figura 12 é apresentado o modelo de dados do banco final do Projeto #2. O resultado é praticamente idêntico ao modelo resultante do Projeto #1. Entretanto, desafios semelhantes ao do Projeto #1 também foram vencidos como por exemplo, ajustes de classe de veículos em relação ao padrão adotado pelo DNIT.

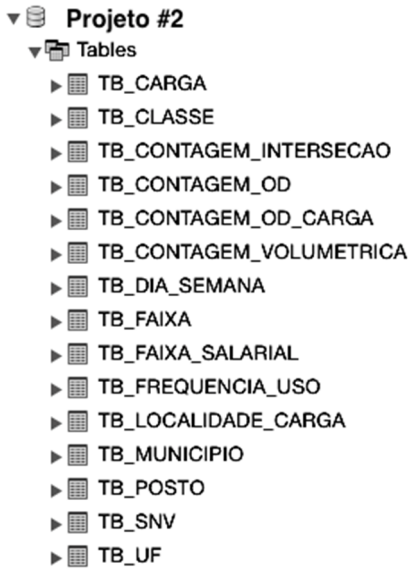

Figura 12: Modelo de dados final Projeto #2, resultante do Processo de Consolidação

Ao fim do processo, um banco de dados relacional foi gerado e importado para um Sistema de Gerenciamento de Banco de Dados (SGBD), no caso deste trabalho foi usado MySQL, como já mencionado. O procedimento para verificação da consistência dos dados importados foi o mesmo empregado no Projeto #1. Também não foram encontrados dados diferentes das planilhas originais. Na Tabela 3 são apresentados os valores das amostras utilizadas.

Tabela 3: Número de amostras utilizada por tipo de pesquisa no Projeto #2.

Tipo de pesquisa	Qtd. de registros	Tamanho da amostra
Volumétrica	191568	384
Origem-Destino (OD)	19826	377

5. CONCLUSÕES

Este artigo apresentou os resultados de uma pesquisa-ação que teve como finalidade o desenvolvimento de um processo de consolidação de dados para pesquisas de tráfego no Departamento Nacional de Infraestrutura de Transportes (DNIT). O processo foi aplicado em oito projetos de contagens de tráfego do órgão para transformar planilhas eletrônicas, documentos de textos e até mesmo figuras, em bases de dados relacionais.

Este trabalho apresentou a aplicação do processo em dois projetos de contagens de tráfego, aqui denominados de Projeto #1 e Projeto #2. Foram recebidos mais de 600 arquivos digitais, entre planilhas, arquivos de texto e fotos. O processo que envolveu análises, modelagens, limpezas, transformações e verificações dos dados, em todos esses arquivos, permitiu a criação de bancos de dados relacionais.

O artigo evidenciou o sucesso da criação e execução do processo de consolidação. O mesmo apresenta rigor científico na busca por soluções e na construção da solução com etapas claras que podem ser auditadas e repetidas. Também fica clara a contribuição do trabalho para a área, com a apresentação e aplicação de um processo de consolidação baseado em técnicas de *data warehouse* e ETL, objetivando resgatar dados históricos que com o passar dos anos, poderiam se perder ou ficar inacessíveis. Acredita-se que este processo pode ser replicado pelas instituições e pesquisadores que utilizam dados de tráfego, auxiliando assim diversos estudos envolvendo, por exemplo, crescimento de tráfego e índices de sazonalidade.

AGRADECIMENTOS

O autor Glaydston Mattos Ribeiro agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Processo No. 307835/2017-0) e a Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ (Processo No. 203.231/2017) pelo suporte financeiro. Os autores ainda agradecem ao Departamento Nacional de Infraestrutura de Transportes – DNIT por todo o apoio fornecido.

REFERÊNCIAS

Altrichter, H., Zuber-Skerritt, O., Kemmis, S. e McTaggart, R. (2002) The concept of action research. *The Learning Organization*, v. 9, n. 3, p. 125-131. DOI:10.1108/09696470210428840

Bitar, N. (2016) *Big Data analytics in transportation networks using the NPMRDS* (PhD Dissertation for Master of Science in Data Science and Analytics). University of Oklahoma, Oklahoma.

Casa Civil. Sistema Nacional de Viação - SNV (2011). *Pub. L. No. 12.379.* Disponível em: http://www.dnit.gov.br/download/sistema-nacional-de-viacao/pnv-lei-5.917/lei-12379-snv.pdf

Casters, M.; R. Bouman e J. van Dongen (2010) *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*. (1º ed). John Wiley & Sons.

Codd, E. (1970) A relational model of data for large shared data banks. *Communications of the ACM*, v. 13, n. 6, p. 377-387. DOI:10.1145/362384.362685

Dhakal, A. (2014) *Data Warehouse Architecture*. *Data Science Central*. Disponível em: http://www.datasciencecentral.com/profiles/blogs/data-warehouse-architecture. Acesso em 13 de junho de 2017.

DNIT (2006) *Manual de Estudos de Tráfego*. Ministério dos Transportes. Disponível em: http://ipr.dnit.gov.br/normas-e-manuais/manuais/documentos/723_manual_estudos_trafego.pdf

DNIT (2017) *Institucional do DNIT*. DNIT. Disponível em://www.dnit.gov.br/acesso-a-informacao/institucional. Acesso em: 13 de junho de 2017

Espadas, J.; D. Romero; D. Concha e A. Molina (2008) Using the Zachman Framework to Achieve Enterprise Integration Based-on Business Process Driven Modelling. In: R. Meersman; Z. Tari e P. Herrero (Eds), *On the Move to Meaningful Internet Systems: OTM 2008 Workshops* (p. 283–293). Springer Berlin Heidelberg.

Herbert A; K. Stephen; M. Robin e Z. S. Ortrun (2002) The concept of action research. *The Learning Organization*, v. *9,* n. 3, 125–131. DOI:10.1108/09696470210428840

Inmon, W. H.; e R. D. Hackathorn (1994) *Using the Data Warehouse*. (1º ed). Wiley Computer Publishing, United States.

Jarke, M., Lenzerini, M., Vassiliou, Y., e Vassiliadis, P. (2013) *Fundamentals of Data Warehouses* (2 ed.). Springer Science & Business Media, New York, USA.

Júnior, L.; C. Abramides; L. Perim e A. Nunes (2016) A retomada da contagem de tráfego rodoviário e atribuição dos componentes do Plano Nacional de Contagem de Tráfego. *Anais do XXX Congresso ANPET*. Apresentado em XXX - ANPET, Rio de Janeiro.

Kakish, K., e Kraft, T. A. (2012) ETL evolution for real-time data warehousing. *Proceedings of the Conference on Information Systems Applied Research*, CONISAR, New Orleans Louisiana, USA, v. 5, n. 2214, p. 1-12.

Kimball, R., e Ross, M. (2013) *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3 ed.). John Wiley & Sons, USA.

Martinovic, L. V. P.; J. O. N. Ferreira; N. E. S. Peixoto e A. P. Fonseca (2016) Transporte Informal de Passageiros: A percepção da comunidade acadêmica da Universidade de Brasília. *Anais do XXX Congresso ANPET*. Apresentado em XXX - ANPET, Rio de Janeiro.

Merriam-Webster. (2018) *Database*. *Merriam-Webster*. Merriam-Webster Inc., Springfield, MA.

Moalla, I.; A. Nabli; L. Bouzguenda e M. Hammami (2017) Data warehouse design approaches from social media: review and comparison. *Social Network Analysis and Mining*, v. 7, n. 1, p. 5. DOI:10.1007/s13278-017-0423-8

Mukherjee, R. e P. Kar (2017) A Comparative Review of Data Warehousing ETL Tools with New Trends and Industry Insight. *Proceedings IEEE 7th International Advance Computing Conference*, p. 943–948. DOI: 10.1109/IACC.2017.0192

MySQL, A. B. (2001) *MySQL reference manual*.

Omair, A. (2014) Sample size estimation and sampling techniques for selecting a representative sample. Journal of Health Specialties, v. 2, n. 4, p. 142–147. DOI:10.4103/1658-600X.142783

Prema, A., e Pethalakshmi, A. (2013) Novel approach in ETL. *Proceedings 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, ICPRIM, Salem, India. v. 1, p. 429–434. DOI:10.1109/ICPRIME.2013.6496515

Tao, T., Parker, S., e Ran, B. (2015) Large Scale Intelligent Transportation System Traffic Detector Data Archiving. *Proceedings of the 15th COTA International Conference of Transportation Professionals,* CITP, Beijing, China. *v.* 1, p. 431-442. DOI: 10.1061/9780784479292.039

Reason, P. e H. Bradbury (2001) *Handbook of action research : participative inquiry and practice*. (1º ed). London, Thousand Oaks.

Sun, R. (2011) *Development of a knowledge base for low-volume roads using a geographic information system* (Master thesis in Civil Engineering). The University of Melbourne, Austrália.

Titirisca, A. (2013) ETL as a Necessity for Business Architectures. *Database Systems Journal*, v. 4, n. 2, p. 3-12.

Von Brown, J.; M. Martello e R. R. Souleyrette (2011) *Minnesota Department of Transportation Traffic Safety Analysis Software State of the Art* (Publication No. MN/RC 2011-10). p. 88. Minnesota Department of Transportation, United States.

Zawodny, J. D. e D. J. Balling (2008) *High Performance MySQL: Optimization, Backups, Replication, Load Balancing & More*. (2º ed). O'Reilly Media, United States.

Zhao, Y., Sandara, M., Huang, S., Sadek, A., George, T., e Hutchins, A. M. (2011) Intelligent Transportation System Data Warehouses and their Applications. *Proceedings 13th International Conference on Enterprise Information Systems*, ICEIS, Beijing, China. v. 1, p. 343-347. DOI:10.5220/0003431503430347