



UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA

2009.2

**TETL: UMA FERRAMENTA DE ETL PARA TRAJETÓRIAS
DE OBJETOS MÓVEIS**

Aluna: Maria Carolina Torres da Silva
Orientadora: Valéria Cesário Times

(mcts@cin.ufpe.br)
(vct@cin.ufpe.br)

Recife, dezembro de 2009

Universidade Federal de Pernambuco
Centro de Informática

Maria Carolina Torres da Silva

**TETL: UMA FERRAMENTA DE ETL PARA TRAJETÓRIAS
DE OBJETOS MÓVEIS**

*Trabalho apresentado ao Programa de Graduação em
Ciência da Computação do Centro de Informática da
Universidade Federal de Pernambuco como requisito parcial
para obtenção do grau de Bacharel em Ciência da
Computação.*

Orientadora: Prof. Valéria Cesário Times

Recife
2009

“Quem acredita, sempre alcança...”

Renato Russo

***A meus pais, pelo apoio, carinho e
amor que sempre me dedicaram.***

Agradecimentos

Primeiramente, gostaria de agradecer a Deus, que me abençoou com tantas coisas boas: minha família, meus amigos, uma vida repleta de paz e saúde, sempre se mostrando presente em todos os momentos, desde os mais alegres aos mais difíceis.

A minha família, por tudo que fizeram por mim durante a minha caminhada, que ainda está em seu início. Não me deixaram faltar nada, sempre me motivando e ajudando, com muita paciência e carinho. Em especial aos meus pais que me ensinaram o verdadeiro sentido das palavras AMOR, AMIZADE e FAMÍLIA; são eles os responsáveis por eu ser quem sou!

A minha orientadora, Valéria Times, que sempre esteve disposta a me ajudar e ouvir quando precisei durante toda a minha graduação, seja apenas como minha professora ou minha orientadora. Pelos vários momentos em que soube me escutar e me dar sábios conselhos, no lugar de broncas e reclamações, mesmo que essas fossem bem mais apropriadas. Por tudo isso, sei que hoje tenho uma nova e boa amiga!

A meus amigos, conquistados antes ou durante a graduação, que sempre estiveram ao meu lado, incentivando, ajudando e crescendo junto comigo. Aos da graduação, além do meu agradecimento, envio também minhas felicitações aos que também estão terminando a graduação e por terem vencido mais esse obstáculo junto comigo; desejo a eles toda sorte e sucesso do mundo.

A todos, o meu eterno e sincero agradecimento!

Resumo

O crescente acesso a tecnologias de dispositivos móveis, como telefone celular, GPS e RFID, tem resultado na geração de enormes quantidades de dados localizados no tempo e no espaço, chamados de trajetórias de objetos móveis. Esse tipo de dado desperta interesse de várias áreas do conhecimento humano, como por exemplo, rastreamento de objetos de valor e monitoramento de criminosos.

Esses dados podem gerar muita informação se análises multidimensionais pudessem ser realizadas sobre eles. Porém, para fornecer suporte a esse tipo de análise, a criação de um *Data Warehouse* é necessária. A especificação e implementação de um *Data Warehouse* de trajetória não é uma atividade simples e ainda é uma área de pesquisa em aberto, por possuir vários componentes ainda sem definição.

Uma dos componentes críticos nesse processo é a realização da etapa de ETL, na qual os dados transacionais são extraídos, transformados e agregados segundo o modelo de dados multidimensional adotado. Atualmente, não existe nenhuma ferramenta de ETL que utilize dados de GPS como entrada, desperdiçando assim, a informação contida nesse tipo de dado.

Para resolver esse problema, o trabalho descrito neste documento propõe uma ferramenta de ETL para extração de informação a partir de dados de GPS. Define também o processo pelo qual essa ferramenta realizará as fases de ETL e um modelo de dados multidimensional diferenciado que considera medidas previamente definidas, dimensões configuráveis e representação de dados oriundos de GPS. O objetivo desse modelo é resolver a falta de padronização entre os modelos propostos por outros autores.

Palavras-chave: trajetória, objetos móveis, Data Warehouse, ETL, GPS

Lista de Figuras

1.1: Metodologia de trabalho proposto.....	3
2.1: Caminho espaço-temporal de objeto móvel e suas trajetórias ([SPAC08]).....	7
2.2: Trajetória com momentos de parada e de movimento	8
2.3: Semântica de uma trajetória ([ABKMMV07])	9
2.4: Fases de um Data Warehouse de Trajetórias ([MARK08])	11
2.5: Modelagem baseada em coordenadas espaciais (adaptação de [OORR07]).....	13
2.6: Modelagem baseada em coordenadas espaciais (adaptação de [BRAZ07])	13
2.7: Modelagem baseada nas informações dos objetos móveis ([MARK08])	14
2.8: Fases de criação de um Data Warehouse ([KAI]).....	16
2.9: Versões dos operadores de agregação ([BDH08]).	19
3.1: Modelo de dados adotado por TETL	22
3.2: Exemplo de trajetória de um objeto móvel e suas observações	23
3.3: Exemplo de mensagem NMEA RMC (adaptação de [NMEA09]).....	26
3.4: Etapas do processo de ETL na ferramenta TETL	28
3.5: Arquitetura do protótipo da ferramenta TETL	31
4.1: Cenário do estudo de caso utilizando TETL	34
4.2: Trecho de arquivo de GPS utilizado como entrada no estudo de caso	35
4.3: Interface com o usuário do protótipo da ferramenta TETL	36
4.4: Trecho da tabela temporária.....	36
4.5: Trecho das tabelas de dimensão	37
4.6: Trecho da tabela de fatos.....	38
5.1: Proposta de reconstituição de trajetórias ([BRAZ07])	40
5.2: Exemplo de outlier em uma trajetória (adaptação de [LHL08])	41
A.1: Modelagem baseada em coordenadas espaciais (adaptação de [OORR07]).....	45
A.2: Modelagem baseada em coordenadas espaciais (adaptação de [BRAZ07])	46
A.3: Modelagem baseada nas informações dos objetos móveis ([MARK08])	46
B.1: Etapas do processo de ETL na ferramenta TETL.....	47

Lista de Quadros

2.1: Análise comparativa dos modelos multidimensionais	15
3.1: Análise comparativa dos modelos multidimensionais existentes versus o proposto	25
3.2: Estrutura da mensagem NMEA RMC [NMEA09].	27

Sumário

Capítulo 1 – Considerações Iniciais	1
1.1 Introdução	1
1.2 Motivação	1
1.3 Objetivos	2
1.4 Metodologia de trabalho	3
1.5 Estrutura do documento	3
Capítulo 2 – Fundamentação Teórica	5
2.1 Introdução	5
2.2 Contextualização	6
2.3 Trajetória	7
2.3.1 Propriedades	7
2.3.2 Semântica	8
2.4 Data Warehouse de Trajetória	10
2.4.1 Modelos de dados multidimensionais	12
2.4.1.1 Modelo baseado em coordenadas espaciais	12
2.4.1.2 Modelo baseado nas informações dos objetos móveis	14
2.4.1.3 Estudo comparativo entre os modelos de dados existentes	15
2.4.2 Processo de ETL	16
2.4.3 OLAP	18
2.5 Conclusões	19
Capítulo 3 – Especificação e implementação da ferramenta TETL	20
3.1 Introdução	20
3.2 A ferramenta TETL	21
3.2.1 Modelo de dados multidimensional proposto	22
3.2.1.1 Análise comparativa entre os modelos existentes e o proposto	24
3.2.2 Fonte e formato dos dados de entrada	26
3.2.3 Etapas do processo de ETL da ferramenta TETL	27
3.3 Protótipo da ferramenta TETL	29
3.3.1 Arquitetura	30
3.3.2 Tecnologias envolvidas	31
3.4 Conclusões	32

Capítulo 4 – Estudo de caso utilizando TETL.....	33
4.1 Introdução	33
4.2 Estudo de caso analisado	33
4.3 TETL em funcionamento.....	35
4.3.1 Interface com o usuário	35
4.3.2 Execução do processo.....	36
4.4 Conclusões	38
Capítulo 5 – Considerações Finais	39
5.1. Introdução	39
5.2. Principais contribuições	39
5.3. Trabalhos futuros	40
Referências	42
Anexos.....	45
Anexo A – Imagens dos modelos de dados analisados	45
Anexo B – Imagem do processo de ETL proposto.....	47

Capítulo 1

Considerações Iniciais

1.1 Introdução

Este capítulo introduz alguns aspectos relevantes sobre o presente trabalho. Inicialmente, na seção 1.2, é exposta a motivação para a realização do mesmo. Na seção 1.3, são encontrados os objetivos a serem alcançados após a elaboração desse trabalho. A metodologia utilizada durante o desenvolvimento do estudo descrito neste documento é apresentada na seção 1.4. Por fim, uma breve descrição da estrutura do documento é dada na seção 1.5.

1.2 Motivação

O recente aumento em tecnologias de dispositivos móveis, como telefone celular, GPS e RFID, tem resultado na geração de enormes quantidades de dados brutos localizados no tempo e no espaço. Estes dados são chamados de trajetórias de objetos móveis [ABKMMV07]. Eles são um novo tipo de dado que está despertando o interesse dos pesquisadores, uma vez que podem auxiliar na solução de problemas reais e atuais em diversas áreas do conhecimento humano, como por exemplo, tráfego urbano, desastres naturais como furacões, migração de aves, monitoramento de criminosos, e outras dezenas de aplicações que utilizam dados localizados no tempo e no espaço.

Um conjunto de objetos móveis é tipicamente armazenado em tabelas relacionais nas quais os registros contêm valores da trajetória $[(X_1, Y_1, T_1), (X_2, Y_2, T_2), \dots, (X_m, Y_m, T_m)]$, que representam o movimento do objeto móvel como uma sequência de posições (X e Y) nos tempos $T_1, T_2, \dots, T_m$ [BDH08].

As pesquisas sobre o desenvolvimento de bases de dados espaço-temporais já estão bastante avançadas e consolidadas, porém ainda há muito a ser pesquisado para que análises multidimensionais em um nível mais abstrato possam ser realizadas sobre trajetórias de objetos móveis. Para fornecer suporte a esse tipo de análise, a solução mais indicada em qualquer domínio é a combinação entre as tecnologias de *Data Warehouse* e as ferramentas OLAP.

O desenvolvimento de um *Data Warehouse* não convencional, como o de trajetórias, não é simples de se especificar e implementar. Muitas pesquisas precisam ser realizadas para que definições sejam formalizadas e ferramentas adequadas sejam disponibilizadas no mercado e na academia. Nesse trabalho, alguns estudos sobre a criação de *Data Warehouse* de trajetórias e a realização de OLAP em trajetórias são descritos.

Uma das etapas mais críticas na implementação de qualquer *Data Warehouse*, seja ele convencional ou não, é a de ETL. É nessa fase que os dados operacionais, que podem estar armazenados de diversas formas (bases de dados relacionais ou objeto relacionais, arquivos texto, XML, etc.), são extraídos, transformados e agregados respeitando as regras de negócio da aplicação, e só posteriormente colocados à disposição no *Data Warehouse* para execução das consultas dos usuários finais.

As ferramentas de ETL atuais normalmente oferecem recursos para extração a partir de documentos de texto, como já foi dito anteriormente. Contudo, documentos de saída de GPS, que trazem informações sobre as trajetórias percorridas pelo objeto móvel, possuem uma formatação única, para os quais as ferramentas de ETL disponíveis no mercado e na academia não provêem suporte. A especificação e implementação de um conjunto de funcionalidades básicas para uma ferramenta de ETL capaz de prover suporte a este conjunto de informações é a motivação principal para a realização deste trabalho.

1.3 Objetivos

O trabalho de pesquisa aqui proposto tem por objetivo especificar componentes que possibilitem a manipulação de dados multidimensionais extraídos de arquivos GPS por ferramentas de ETL, para que se torne viável a criação de um *Data Warehouse* de trajetórias de objetos móveis. É essencial incluir este formato no processo de ETL, uma vez que, diante da crescente utilização de receptores de sinais de GPS, muita informação é gerada, mas não é aproveitada. Isso ocorre devido ao fato de que a manipulação de trajetórias em um ambiente de decisão ainda não é realizada de forma eficiente, segura e automatizada.

Dessa forma, os objetivos gerais deste trabalho são:

- Análise e aplicação dos conceitos de ETL e *Data Warehouse* de dados convencionais para uniformização de terminologias e avaliação da aplicabilidade deles no contexto desse trabalho: trajetórias de objetos móveis;
- Incluir, no processo de ETL, arquivos de GPS como entrada de dados, com o objetivo de gerar conteúdo para o desenvolvimento de um *Data Warehouse* de Trajetórias.
- A especificação e implementação de um conjunto de funcionalidades básicas para uma ferramenta de ETL capaz de prover suporte ao conjunto de dados de posicionamento que são atualmente disponíveis e gerados em abundância, e não são adequadamente utilizados.

1.4 Metodologia de trabalho

A metodologia utilizada para a realização do presente trabalho está sistematizada e apresentada na Figura 1.1. Inicialmente é realizada a revisão de literatura, etapa em que há a análise do estado da arte do domínio de estudo envolvido. Após essa etapa inicial, define-se o escopo da problemática a ser abordada. Partindo do escopo já definido, há a especificação e implementação do trabalho descrito neste documento. Em seguida, o trabalho realizado é validado por meio do desenvolvimento de um estudo de caso.

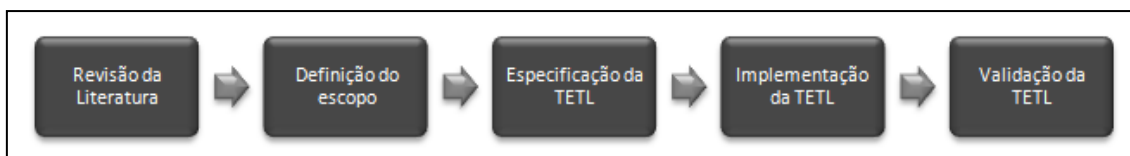


Figura 1.1: Metodologia de trabalho proposto

1.5 Estrutura do documento

A seguir será descrita a distribuição dos tópicos ao longo dos capítulos deste documento.

- **Capítulo 2 – Fundamentação Teórica:** este capítulo introduz as principais definições envolvidas no domínio abordado. Conceitos, modelos e estruturas são exibidos nesse capítulo na intenção de serem

utilizados eventualmente durante o desenvolvimento do trabalho aqui descrito;

- **Capítulo 3 – Especificação e implementação da ferramenta TETL:** este capítulo apresenta aspectos envolvidos nas fases de especificação e implementação da ferramenta proposta. Dentre as muitas informações dispostas nesse capítulo, estão: a arquitetura utilizada, os requisitos do software, os componentes desenvolvidos, entre outras;
- **Capítulo 4 – Estudo de caso:** capítulo responsável pela apresentação de um estudo de caso utilizando a ferramenta especificada e apresentada no capítulo anterior, demonstrando seu funcionamento e sua utilidade;
- **Capítulo 5 – Considerações Finais:** neste capítulo, se encontram as principais conclusões retiradas desse trabalho. Nele, são descritos os conteúdos de maior importância gerados durante esse estudo, bem como a indicação de trabalhos futuros para o melhoramento da ferramenta de ETL aqui proposta.

Capítulo 2

Fundamentação Teórica

2.1 Introdução

Graças ao crescimento no uso de sensores e de equipamentos de localização, tornou-se técnica e economicamente viável a captura em larga escala de dados acerca das posições de objetos móveis. Isso proporcionou uma nova perspectiva para inúmeras aplicações baseadas nessas informações, obtidas a partir de pessoas ou aviões munidos de equipamentos GPS, animais que possuam transmissores cujos sinais sejam capturados por satélite, entre tantas outras possibilidades [SPAC08].

Essas informações precisam ser organizadas para que mantenham consistência, gerando conhecimento e servindo como base na tomada de decisões. Para tanto, a utilização de *Data Warehouses* se mostra uma opção interessante, porém essa tecnologia ainda não foi suficientemente investigada para funcionar satisfatoriamente sobre dados de trajetória, precisando ainda de uma fundamentação conceitual e prática. Para aplicar a tecnologia de *Data Warehouse* sobre trajetórias, muitos estudos estão sendo realizados e documentados, e alguns deles são descritos neste capítulo.

O objetivo principal desse capítulo é fornecer ao leitor, conceitos importantes para facilitar o entendimento sobre o trabalho realizado, provendo fundamentos para que a leitura deste documento ocorra da melhor maneira possível. A fim de atingir tal objetivo, na seção 2.2, uma breve contextualização do domínio abordado é descrita. A seção 2.3 objetiva a apresentação de alguns conceitos fundamentais acerca do objeto de estudo principal desse trabalho: a trajetória. Já a seção 2.4 inclui fundamentos, problemas e tendências sobre o tópico mais importante desse trabalho: o desenvolvimento de *Data Warehouse* de Trajetória. Finalmente, a seção 2.5 exhibe algumas conclusões obtidas a partir das análises sobre trabalhos correlatos que são apresentadas nas seções anteriores.

2.2 Contextualização

O uso de aparelhos de localização como telefones celulares e equipamentos de GPS teve um notável crescimento nos últimos anos, permitindo acesso a um imenso conjunto de dados alocados no tempo e no espaço [MARK08], caracterizando o que se denomina de dados espaço-temporais. A aquisição destes dados resultou na geração de outro tipo de dado: o de trajetória, que pode ser definido como a evolução da posição de um objeto que está se movendo no espaço durante um intervalo de tempo a fim de alcançar um dado objetivo [SPAC08].

São inúmeras as aplicações que podem ser beneficiadas a partir da representação e manipulação dos dados de trajetória. Alguns exemplos são: controle de tráfego urbano ou aéreo, migrações de animais, rastreamento de objetos móveis, monitoramento de criminosos, entre outros. As aplicações podem ser em tempo real ou não, e os dados de trajetória podem ser utilizados em Sistemas de Suporte à Decisão nas aplicações citadas.

Uma das tecnologias amplamente utilizadas no contexto de Sistemas de Suporte à Decisão é o *Data Warehouse* (DW) que pode ser definido como uma base de dados integrados, orientados por assunto, não volátil e variante no tempo [INM05]. Normalmente, os dados armazenados em um DW servem de base para a aplicação da tecnologia OLAP (*On-Line Analytical Processing*). OLAP é uma categoria de software específica para realizar processamento analítico nos dados de um DW [CHDAY97]. Porém, ainda não existe um consenso na literatura de BD sobre a definição de uma arquitetura, de modelos de dados e de ferramentas de consulta para um DW de trajetórias.

A trajetória, além de ser um tipo de dado espaço-temporal, possui informação semântica, tais como momentos de parada e de movimento que não são facilmente extraídos a partir de sua geometria [SPAC08]. A extração, representação e manipulação desta informação semântica em um DW de trajetórias, aumentam o universo de análises que podem ser realizadas sobre os dados.

Sistemas gerenciadores de bancos de dados convencionais e sistemas gerenciadores de bancos de dados geográficos ou espaço-temporais, em geral, não oferecem o completo suporte para análise e mineração de informações sobre objetos móveis, ou seja, trajetórias [WOL98]. Desta forma, não se tem conhecimento até o momento, sobre a existência de ferramentas comerciais que permitam o gerenciamento,

o cruzamento e a realização de consultas analíticas envolvendo este tipo de dado. Além disso, é desconhecida a existência de uma ferramenta de ETL que extraia informações de trajetórias a partir de dados de GPS.

2.3 Trajetória

Esta seção descreve as principais características de uma trajetória, além de detalhar aspectos importantes relacionados à incorporação de semântica aos dados de trajetórias de objetos móveis.

2.3.1 Propriedades

Os dados de trajetória possuem, além de informações sobre suas posições geográficas no tempo, informações semânticas de seu percurso. Muitos estudos objetivam determinar os conceitos acerca dessas informações, como a formalização do que é uma trajetória, um momento de parada ou um momento de movimento. Após a análise de alguns destes estudos [AATOM09, ABKMMV07, PBKA08], percebe-se que as definições mais aceitas pela academia são as apresentadas por Spaccapietra *et al.* (2008).

Spaccapietra *et al.* (2008) definem uma **trajetória** como sendo o registro da evolução da posição de um objeto que está se movendo no espaço durante um determinado intervalo de tempo a fim de alcançar um dado objetivo. Tal trajetória pode ser basicamente representada por pares compostos de um ponto e uma medida de tempo. Como uma trajetória pode ser vista como um segmento de caminho espaço-temporal, percorrido por um objeto móvel, pode-se concluir que, durante o tempo em que um objeto tem seu movimento observado, ele pode percorrer várias trajetórias, uma após a outra.

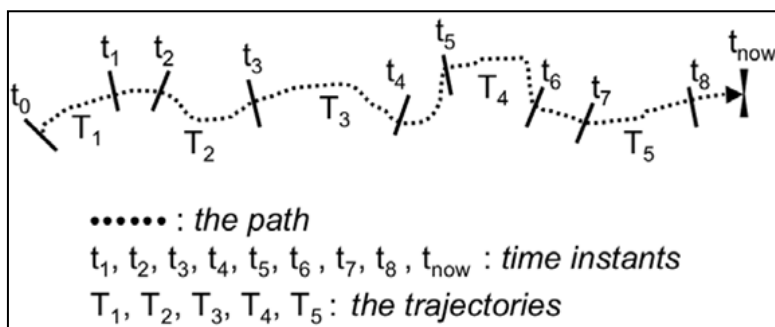


Figura 2.1: Caminho espaço-temporal de objeto móvel e suas trajetórias ([SPAC08])

Na Figura 2.1, pode-se observar a existência de intervalos de tempo em que não há a categorização da ocorrência de pontos de uma trajetória, como entre os tempos t_1 e t_2 . Isso ocorre devido a esse movimento ser irrelevante para a aplicação. Então, observando a Figura 2.1, conclui-se que há quatro intervalos de tempo que soa considerados irrelevantes para uma dada aplicação. Esses intervalos são: $[t_1, t_2]$, $[t_4, t_5]$, $[t_6, t_7]$ e $[t_8, t_{\text{now}}]$.

Spaccapietra *et al.* (2008) também definem o conceito de **momento de parada** durante uma trajetória, que pode ser definido pelas três regras a seguir: (1) o usuário define explicitamente a parte da trajetória ($[T_{\text{inícioParada}}, T_{\text{fimParada}}]$) que representa uma parada; (2) o tempo percorrido no intervalo especificado é um intervalo não vazio e (3) o objeto móvel não se move, ou seu movimento é imperceptível à aplicação. Na Figura 2.2, encontram-se exemplos de momentos de paradas, tais como em P1 e P2.

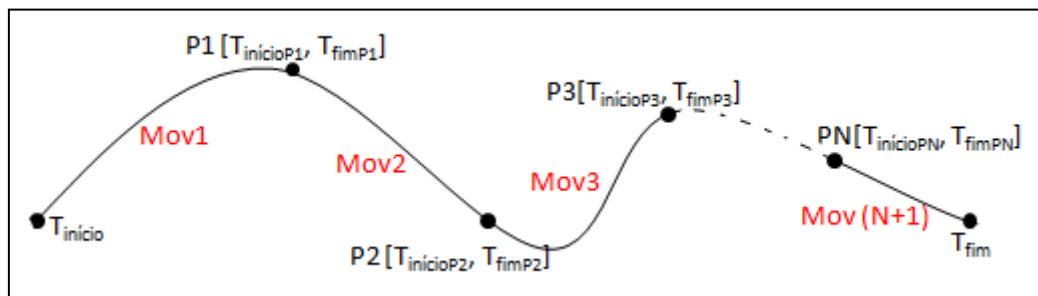


Figura 2.2: Trajetória com momentos de parada e de movimento

Há ainda a conceitualização do que seria um **momento de movimento** em uma trajetória, definido pelas seguintes regras: (1) um movimento é limitado por duas extremidades que representam: duas paradas consecutivas (como pode ser observado em Mov2 ou Mov3 mostrados na Figura 2.2), ou $T_{\text{início}}$ e a primeira parada (denotado por Mov1), ou a última parada e T_{fim} (representado por Mov (N+1) na Figura 2.2), ou ainda $T_{\text{início}}$ e T_{fim} , caso a trajetória em questão não possua qualquer parada durante o seu percurso; (2) o tempo percorrido no intervalo especificado é um intervalo não vazio e (3) o espaço percorrido no intervalo especificado é uma ou mais linhas espaço-temporais.

2.3.2 Semântica

Sistemas gerenciadores de bancos de dados convencionais e sistemas gerenciadores de bancos de dados geográficos ou espaço-temporais, em geral, não oferecem o suporte adequado para a análise e mineração de trajetórias. Além disso, os

protótipos existentes de bancos de dados de objetos móveis se preocupam apenas com as propriedades geométricas destes dados, não provendo, portanto, suporte a integração das trajetórias geradas por sistemas de posicionamento com dados geográficos ou com dados convencionais para agregar algum tipo de semântica [ABKMMV07].

A idéia de agregar semântica às trajetórias de objetos móveis para facilitar a análise e mineração é uma idéia inovadora que foi introduzida por Alvares *et al.* (2007). Os autores criaram um modelo para adicionar semântica às trajetórias, a fim de facilitar e otimizar consultas espaço-temporais, já que trajetórias semânticas ajudariam no desenvolvimento de diversas aplicações de variados domínios.

A trajetória bruta é representada de forma simples, possuindo informações apenas sobre um conjunto de pares (x, y, t) , indicando uma posição no espaço e no tempo. A partir do uso exclusivo dessa representação, torna-se muito difícil encontrar algum padrão semântico novo e interessante. Além disso, às vezes, os padrões descobertos podem não ser significativos em determinados domínios de aplicação. Porém, uma trajetória bruta pode ser enriquecida com diferentes informações semânticas, de acordo com o contexto da aplicação e o problema que o usuário pretende resolver [BOWA08].

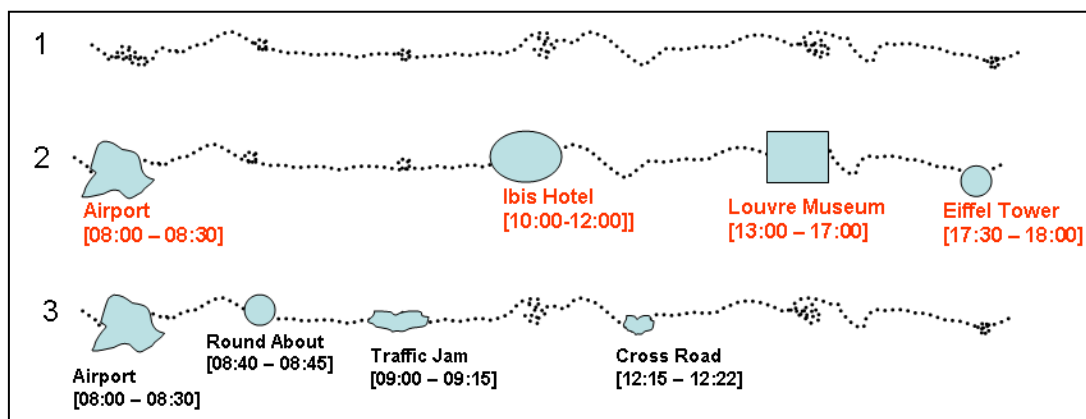


Figura 2.3: Semântica de uma trajetória ([ABKMMV07])

Na Figura 2.3, encontra-se três exemplos de trajetórias, sendo a primeira delas, uma trajetória bruta e as demais, semanticamente enriquecidas para diferentes domínios de aplicação. A trajetória 1 é apenas um conjunto de vários pontos (x, y, t) que compõem uma trajetória sem nenhuma contextualização. Na trajetória 2, a semântica se insere no contexto de uma aplicação de turismo, na qual os aspectos importantes são locais como aeroportos, hotéis e pontos turísticos. Já na terceira trajetória, a semântica está relacionada a uma aplicação de trânsito, na qual as partes interessantes

correspondem a lugares movimentados (aeroportos), rotatórias, cruzamentos, sinaleiras, pardais, etc. Este exemplo comprova que os mesmos dados brutos, após a inserção de semântica, podem ser utilizados em diferentes domínios de aplicação.

Na área de estudo sobre trajetórias semânticas existem muitos temas de pesquisa ainda em aberto. Como exemplos, podem ser citados: o desenvolvimento de algoritmos para encontrar padrões semânticos a partir de trajetórias brutas, a especificação e implementação de linguagens e operadores para agregação semântica em ambientes multidimensionais, além da integração de forma incremental, automática e eficiente de dados de GPS e informação semântica sobre a localização. Porém, a incorporação de semântica a dados de GPS está fora do escopo do trabalho descrito nesse documento.

2.4 *Data Warehouse de Trajetória*

Trajeto rias possuem sua representa  o simples, mas s o dados altamente complexos, j  que envolvem a agrega  o de diferentes tipos de informa  o, como geom trica, geogr fica e de contexto. Trajet rias, por envolverem a dimens o espacial e temporal, requerem agrega  es complexas de tempo e espa o, e a tecnologia que melhor se adequa a esses requisitos   *Data Warehouse*. Essa tecnologia objetiva agregar dados em m ltiplas dimens es, sendo, ent o, de extrema utilidade para fazer an lises avan adas em trajet rias [OORRS07] [PRDVMFNT08].

DW possibilita a an lise de grandes volumes de dados, de eventos atuais e eventos passados, oferecendo suporte para a tomada de decis o e para a previs o de eventos futuros [ELNA03]. Aplica  es desse tipo s o mais aplicadas aos n veis superiores de uma organiza  o, gerencial ou estrat gico, e n o ao operacional. Por isso, a aplica  o dessa tecnologia na  rea de trajet rias   de extremo interesse comercial e acad mico.

O processo de desenvolvimento de um *Data Warehouse* de Trajet rias (DWT) envolve muitas fases, as quais podem ser observadas na Figura 2.4. A etapa (1) envolve a coleta dos dados das trajet rias dos objetos m veis analisados, ou seja, a aquisi  o de suas posi  es no tempo e no espa o que est o representadas pelos pares (x, y, t) . Esses dados precisam ser reconstru dos para a composi  o de uma trajet ria.   nessa etapa que as corre  es necess rias s o realizadas. Ap s a reconstru  o ter sido finalizada, na etapa (2), os dados s o armazenados em uma base de dados de objetos m veis, de

acordo com o modelo de dados do ambiente transacional. Porém, a realização do armazenamento realizado na etapa (2) não é obrigatória, já que os dados do DW podem ser gerados automaticamente a partir dos dados originais que são coletados a partir do posicionamento dos objetos móveis.

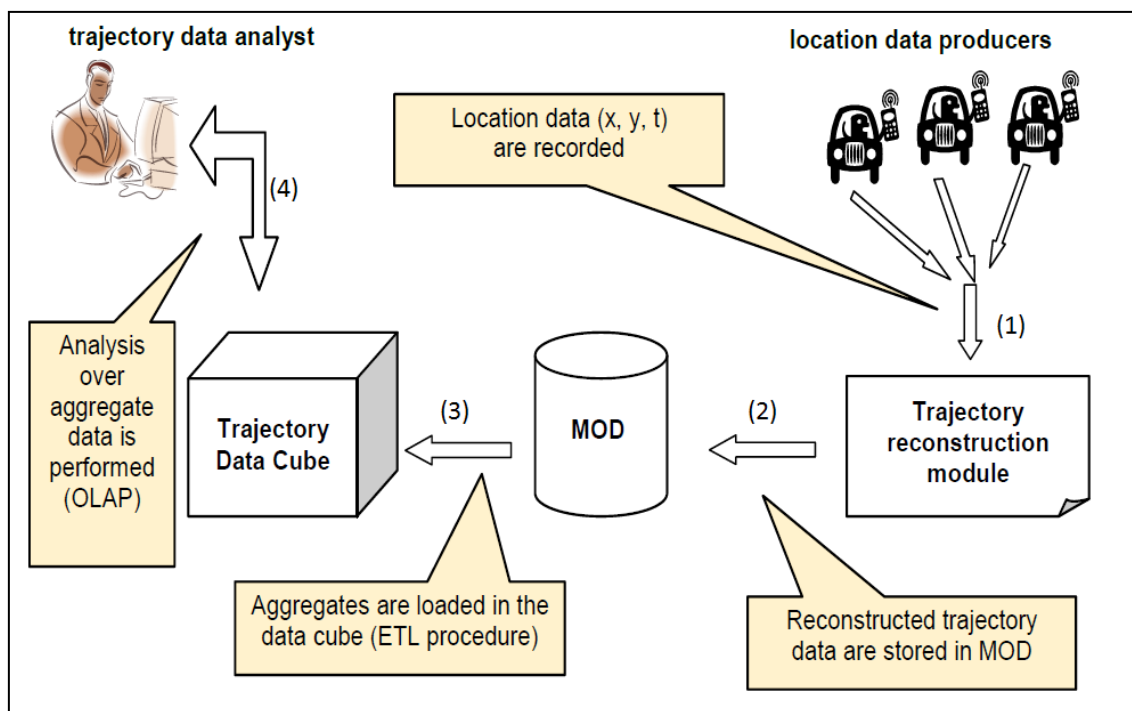


Figura 2.4: Fases de um Data Warehouse de Trajetórias ([MARK08])

O processo de ETL é aplicado na etapa (3). Esse processo é o responsável por realizar as agregações necessárias, respeitando o modelo de dados do *Data Warehouse* no qual os dados ficarão armazenados. Após o DW estar efetivamente povoado com os dados já agregados, os cubos podem ser projetados e utilizados pelos usuários como fonte de informações analíticas por meio do uso de ferramentas OLAP (ver passo passo 4 da Figura 2.4).

Esta seção pretende discutir os principais problemas e tendências sobre o desenvolvimento de um DWT. Na seção 2.4.1, as propostas de modelo de dados para DWT são discutidas, incluindo a especificação de medidas e dimensões. Na seção 2.4.2, abordaremos os problemas envolvidos na etapa de extração, transformação e carga dos dados de um DWT, apresentando algumas propostas já existentes. A seção 2.4.3 objetiva expor os estudos acerca da especificação e do desenvolvimento dos operadores de consulta OLAP utilizados em um DWT.

2.4.1 Modelos de dados multidimensionais

Após análise da literatura sobre o domínio abordado, nota-se que não existe um consenso acerca do modelo de dados multidimensional a ser adotado na implementação de um DWT, sobre o qual os cubos serão gerados e as consultas OLAP realizadas. Porém, há algumas propostas na literatura de banco de dados que especificam e definem um modelo de dados para DWT.

Cada proposta determina um conjunto de medidas e dimensões, define como deve ser feita a hierarquização dentro de cada dimensão, e geralmente, formaliza as agregações a serem implementadas nas aplicações previamente especificadas. Porém, a medida mais importante é denominada de *presença*. Essa medida representa a quantidade de trajetórias em uma célula, que é a representação de um quadrante espacial em um determinado intervalo de tempo.

Essa seção exibe informações sobre alguns modelos de dados para DWT sem considerar a relevância de cada um deles na ordem de apresentação. A seção 2.4.1.1 descreve o modelo de dados multidimensional baseado nas coordenadas espaciais. Na seção seguinte, a 2.4.1.2, uma proposta de modelo de dados multidimensional baseado nas informações dos objetos móveis é descrita. Na última seção, a 2.4.1.3, um estudo comparativo entre os modelos propostos e descritos nas seções anteriores é sumarizado.

2.4.1.1 Modelo baseado em coordenadas espaciais

Orlando *et al.* (2007) e Braz (2007) propõem uma abordagem para a aplicação dos conceitos básicos de um DW clássico no domínio de trajetórias. O objetivo dos autores é definir um modelo de dados eficiente com foco na trajetória em si, sem a preocupação com o objeto móvel responsável pela geração de cada trajetória. Uma sistematização da proposta realizada por Orlando *et al.* (2007) pode ser visualizada na Figura 2.5. Já a proposta feita por Braz (2007), pode ser observada na Figura 2.6. Cada uma destas abordagens é discutida a seguir.

Ao se analisar a Figura 2.5(a), pode-se observar que duas dimensões acerca das posições espaciais (dimX e dimY, referentes às coordenadas X e Y, respectivamente) e uma temporal (dimT) foram especificadas. Uma particularidade desse modelo é o mecanismo de hierarquização aplicado a todas as dimensões, que pode ser visto na

Figura 2.5(b). Nele os níveis utilizados são baseados em intervalos de tempo ou de espaço, dependendo da dimensão.

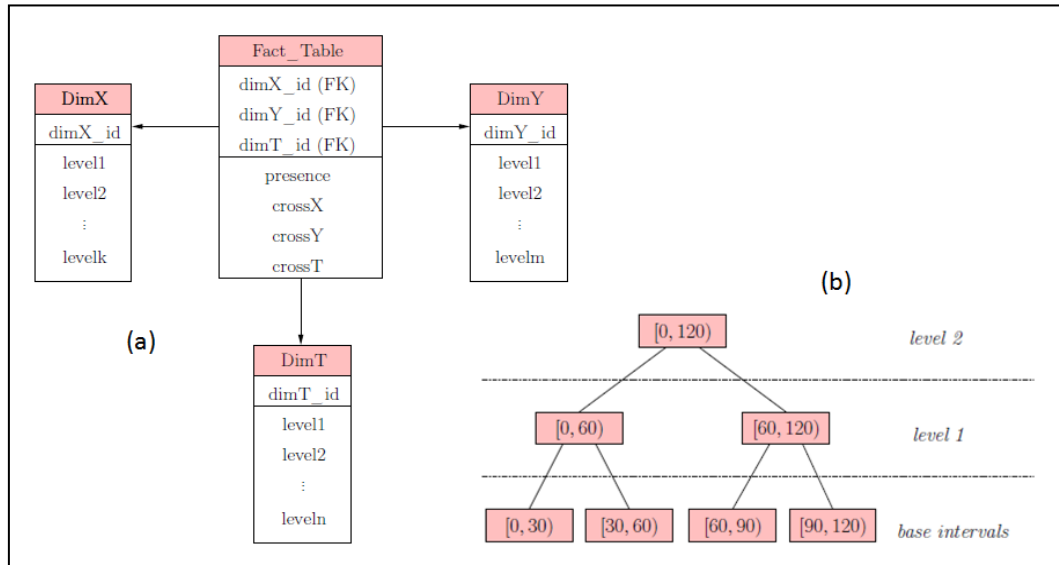


Figura 2.5: Modelagem baseada em coordenadas espaciais (adaptação de [OORR07])

As medidas propostas por Orlando *et al.* (2007) são: a quantidade de trajetórias, a distância percorrida, a velocidade e aceleração máxima, mínima e média.

Braz (2007) propõe uma abordagem semelhante ao modelo de Orlando *et al.* (2007) quanto a definição da hierarquização das dimensões espaciais e da dimensão temporal, as quais são organizadas em intervalos, seguindo o padrão encontrado na Figura 2.5(b). Ele especifica as seguintes medidas para o seu modelo: o número de observações na célula, a quantidade de trajetórias que começam nessa célula, a velocidade média, a quantidade de trajetórias que cruzam algum dos eixos, entre outras que podem ser observadas na tabela de fatos da Figura 2.6.

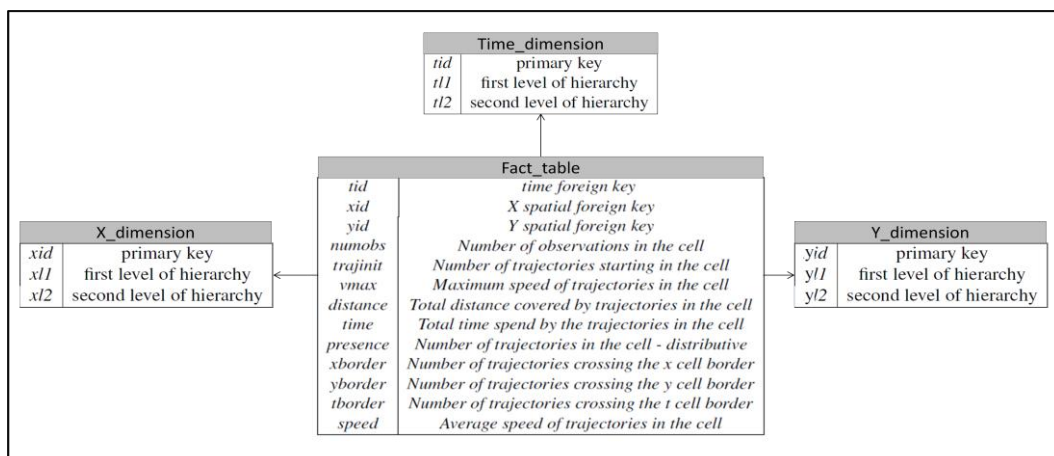


Figura 2.6: Modelagem baseada em coordenadas espaciais (adaptação de [BRAZ07])

Ambas as propostas discutidas nesta seção ([OORR07] e [BRAZ07]) propõem várias medidas, mas apenas a medida *presença* é formalmente definida, incluindo a especificação dos problemas acerca dessa medida durante a fase de carga e cálculo das agregações do processo de ETL. O algoritmo utilizado para o cálculo dessa medida nas duas propostas pode ser encontrado em Braz *et al.*(2007).

2.4.1.2 Modelo baseado nas informações dos objetos móveis

Marketos *et al.* (2008) propõem um framework para criação de DWT. Esse framework é mostrado na Figura 2.4. Ele inclui propostas de projeto e implementação de várias etapas do processo de criação de um DWT, desde a etapa de aquisição dos dados até a formulação das agregações para cada medida, passando pelas etapas de reconstrução das trajetórias e armazenamento em bases de dados relacionais como um passo intermediário.

A modelagem multidimensional proposta por Marketos *et al.*(2008) é exibida na Figura 2.7. A proposta possui três dimensões: (1) Geográfica, que permite definir a hierarquia espacial, no qual o espaço é dividido em pequenas áreas, que para critério de simplificação no artigo, são retangulares e configuráveis pelo usuário (por exemplo: áreas de 10x10 Km²); (2) Tempo, que define a hierarquia temporal, a qual possui uma particularidade pelo uso de um intervalo, sendo também configurável pelo usuário, como a definição do nível mais baixo de granularidade; e (3) Perfil do objeto, contendo informações sobre o objeto em movimento.

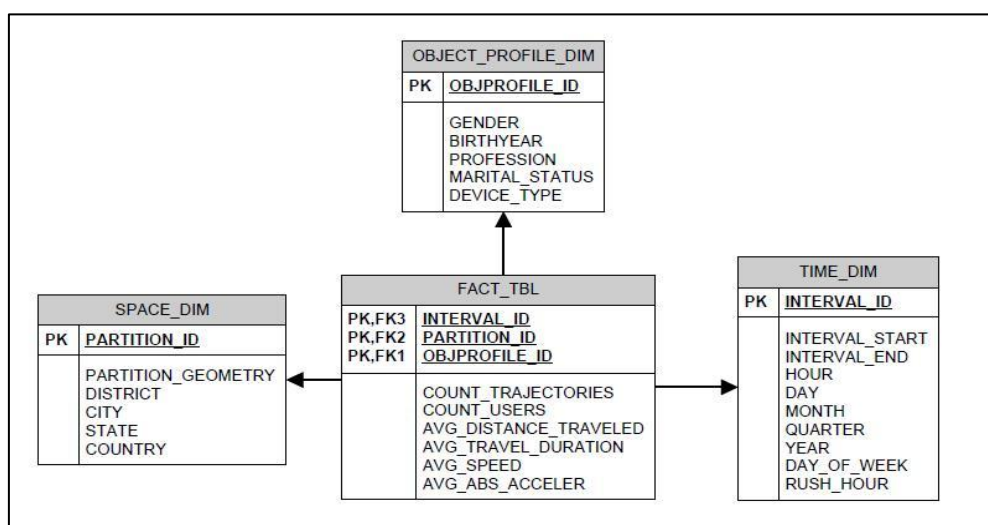


Figura 2.7: Modelagem baseada nas informações dos objetos móveis ([MARK08])

As medidas propostas por Marketos *et al* (2008), que também podem ser visualizadas na Figura 2.7, são: a quantidade de trajetórias, a quantidade de usuários diferentes; a distância e a duração médias dos percursos dos usuários, a velocidade e a aceleração médias durante a realização desses percursos.

No modelo aqui descrito, a medida chamada de quantidade de trajetórias é equivalente à medida *presença*. O cálculo dessa medida é realizado fazendo o uso do mesmo algoritmo utilizado no modelo anterior, especificado por Braz *et al.*(2007).

2.4.1.3 Estudo comparativo entre os modelos de dados existentes

Essa seção realiza uma comparação entre os modelos apresentados nas seções anteriores. O Quadro 2.1 apresenta o resultado da análise comparativa que foi feita entre os modelos descritos anteriormente utilizando os seguintes critérios: (1) representa a preocupação do modelo em representar os objetos móveis envolvidos, ou seja, se o modelo fornece suporte à representação de informações adicionais além dos dados das trajetórias, como, por exemplo, dados sobre os objetos que as geraram; (2) avalia se todas as medidas sugeridas pelo modelo possuem definições, não sendo apenas citadas no texto; (3) considera se as dimensões podem ser configuradas pelo usuário do modelo; (4) analisa se o modelo permite a representação de dados de GPS; e (5) define se alguma aplicação de ETL, que possua como entrada dados de GPS, utiliza o modelo proposto.

Quadro 2.1: Análise comparativa dos modelos multidimensionais

Crítérios	Coordenadas espaciais	Informações dos objetos móveis
Representação de atributos de objetos móveis		✓
Definição das medidas propostas		✓
Dimensões configuráveis	✓	✓
Representação de dados de GPS	✓	
Utilização em uma ferramenta de ETL para dados de GPS		

Além de não existir um modelo de dados padrão para DWT, o estudo comparativo entre as abordagens existentes mostra que dados gerados por GPS não são

considerados como fontes de entrada para ferramentas de ETL, mesmo que existam modelos de dados que permitam representá-los. Ademais, a incorporação de atributos dos objetos móveis no modelo torna a abordagem dependente do domínio da aplicação, uma vez que tais propriedades são específicas e dependem da área de estudo sendo modelada. Por isso, a representação de propriedades dos objetos móveis não é considerada na maioria das abordagens existentes.

2.4.2 Processo de ETL

O processo de ETL, do Inglês *Extract Transform Load* (Extração Transformação Carga), possui como principais funções: (1) a extração de dados de diversos sistemas, (2) a transformação desses dados conforme regras de negócios e (3) a carga efetiva dos dados em *Data Warehouses* ou *Data Marts*. É considerada uma das fases mais críticas da criação de DW.

O maior objetivo de um processo de ETL é disponibilizar os dados de fontes variadas de forma integrada em um formato adequado para ser armazenado no DW. Para que os dados sejam armazenados no DW de forma consistente para a tomada de decisão, é necessário que diversos requisitos sejam atendidos. *Ross and Kimball* (2004) citam como exemplos destes requisitos: (1) atender as necessidades do negócio; (2) atestar a conformidade dos dados fornecidos; (3) identificar o perfil dos dados nas fontes de dados; (4) garantir as restrições de segurança; (5) identificar o nível de integração; (6) definir a periodicidade de disponibilização dos dados; (7) estabelecer os instantes de materialização dos registros em bases de dados temporárias e, por fim; (8) conhecer as linguagens e ferramentas envolvidas na execução do processo de ETL.

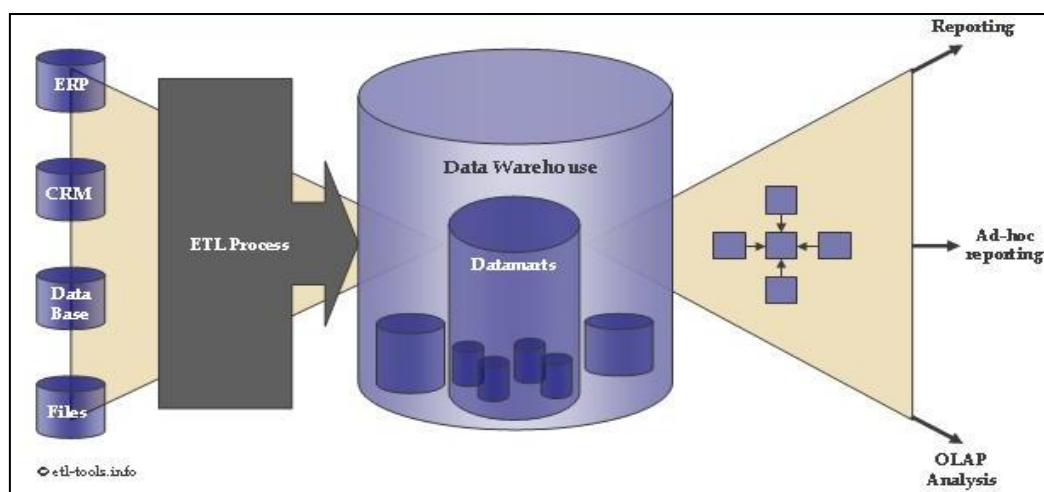


Figura 2.8: Fases de criação de um Data Warehouse ([KAI])

Durante o processo de ETL, muitas fontes de dados podem ser acessadas, as quais podem usar diferentes formas de armazenamento. Na Figura 2.8, pode-se observar como exemplo, que o processo de ETL agrega dados armazenados em bases de dados, que podem estar modelados sob qualquer paradigma (relacionais ou não), arquivos comuns (textos comuns, XML ou de qualquer outra formatação), CRM e quaisquer ferramentas de auxílio utilizadas pela empresa ou organização. É possível observar ainda que o resultado desse processo pode ser um *Data Warehouse* completo, composto por vários *Datamarts*.

A fase de ETL que antecede a criação de um DWT também é crítica e é dependente diretamente do modelo multidimensional adotado e dos requisitos que precisam ser atendidos, assim como ocorre na criação de um *Data Warehouse* convencional.

Como vimos na seção 2.3, não existe um modelo de dados padrão para que possa ser utilizado na modelagem multidimensional de um DWT. Os autores são divergentes sobre as medidas e dimensões que devem ser utilizadas, porém todos os autores fazem uso de uma característica particular associada às dimensões espaciais e temporais. Essa característica diz respeito aos níveis da hierarquia sob a qual elas devem ser montadas.

Essa particularidade presente nos modelos apresentados pode ser analisada com mais clareza na Figura 2.5(b). Os níveis de hierarquia são definidos por intervalos e não por valores únicos como normalmente ocorre em DW convencionais. Esse requisito exige um maior cuidado na implementação das agregações.

Um requisito importante que influencia na complexidade do sistema responsável pelo processo de ETL é o formato e a frequência com que os dados de entrada serão disponibilizados, levando em consideração se o sistema que faz uso dessas informações em tempo real ou não. Os dados podem não estar numa mesma unidade de medida, exigindo a realização de operações de conversão e a entrada de dados pode não ser feita de forma regular. Quando se trata de um DWT, a determinação desses requisitos também é de extrema importância para a criação de um processo ou de uma ferramenta de ETL.

Além disso, as medidas a serem calculadas durante o processo de ETL de um DWT variam de acordo com o modelo adotado, como visto na seção 2.3. Logo, as

fórmulas para seus cálculos são dependentes do modelo utilizado, das medidas adotadas e das regras de negócio que precisam ser respeitadas. Por isso esse trabalho especifica um processo de ETL para trajetórias geradas por GPS, incluindo um modelo de dados multidimensional utilizado pela ferramenta TETL que implementa o processo proposto.

2.4.3 OLAP

A essência de um DW é prover o máximo de flexibilidade aos usuários para que eles possam sumarizar e analisar os dados analiticamente. Isso é realizado por meio da definição de um cubo de dados (com suas dimensões, hierarquias, níveis e fatos relacionados) e a aplicação de operadores OLAP. Esses operadores são responsáveis por: (1) agregar ou desagregar informações por cada dimensão, utilizando-se dos operadores de *drill-down* e *roll-up*; (2) selecionar uma parte específica de uma dimensão, por meio do uso dos operadores de *slice*; (3) selecionar entre duas ou mais dimensões com o uso de operadores de *dice*; e (4) prover apresentações alternativas dos dados, facilitando a visualização dos dados pelo usuário por meio de operadores de *pivot* [BOORRS06].

A motivação para o desenvolvimento de DWT é transformar dados de trajetórias brutas em informação que possa ser utilizada em sistemas de suporte à decisão [BDH08]. Mas para isso ser possível, uma linguagem de consulta precisa ser especificada e implementada, ou seja, é necessário o desenvolvimento de uma linguagem OLAP para análise multidimensional de trajetórias.

Na literatura, é possível encontrar autores que defendam a idéia de que a linguagem OLAP para trajetórias deve ser uma especialização da linguagem OLAP para *Data Warehouses* espaciais, conhecida como linguagem SOLAP. Eles afirmam que por meio da introdução de um nível de abstração mais alto no topo dos dados espaço-temporais resolve-se o problema. Porém, OLAP de trajetórias requer operadores adicionais aos já existentes em SOLAP, além da especificação de novas funções sobre um conjunto de trajetórias [BOORRS06], ou seja, trajetórias precisam de seu próprio conjunto de operadores semânticos para permitir que as análises multidimensionais possam ser realizadas. Até o presente momento, não existe muitos trabalhos relevantes na literatura que especifique operadores OLAP para dados de trajetórias, mas um deles [BDH08] merece uma atenção maior e por isso, ele é detalhado a seguir.

Baltzer *et al.* (2008) definem um operador de agregação OLAP para trajetórias, denominado de GROUP_TRAJECTORIES. Esse operador pode agregar as trajetórias de três formas: (1) por sobreposição, (2) por interseção, e (3) pela combinação das duas características anteriores.

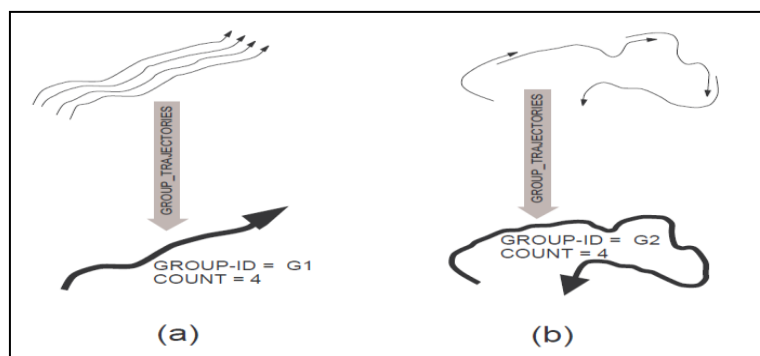


Figura 2.9: Versões dos operadores de agregação ([BDH08]).

Na Figura 2.9, pode-se observar o resultado da aplicação das duas primeiras formas do operador proposto por Baltzer *et al.* (2008). A Figura 2.9 (a) ilustra a aplicação do operador que realiza a agregação por interseção, enquanto a Figura 2.9 (b) exibe o resultado obtido com a aplicação da agregação por sobreposição.

Após a criação do DWT e o cubo de dados já ter sido devidamente processado, os dados estão disponíveis como fonte de consultas analíticas por meio da aplicação de operadores OLAP. Esta é uma área de pesquisa em aberto, mas que foge do escopo do presente trabalho.

2.5 Conclusões

O desenvolvimento de DWT ainda é uma área de pesquisa que precisa de muito investimento e estudo. Muitos dados são gerados devido ao crescente uso de aparelhos com sistemas de localização. Porém, eles são pouco usados na geração de informação agregada e de apoio na tomada de decisões.

Mesmo sendo um domínio de interesse de várias áreas, como controle de trânsito urbano, monitoramento de criminosos ou de catástrofes naturais, rastreamento de frotas ou de transporte de valores, dentre outras, não existe ainda um consenso sobre os componentes que devem existir em um sistema de DWT ou em um módulo de dados para DWT. Dentre os vários componentes ainda não definidos, a especificação de uma ferramenta de ETL para DWT é uma questão fundamental e ainda em aberto. Essa será a questão abordada no próximo capítulo.

Capítulo 3

Especificação e implementação da ferramenta TETL

3.1 Introdução

Como já foi explicado no capítulo anterior, um dos maiores problemas no desenvolvimento de qualquer tipo de DW é a etapa de ETL. Com o objetivo de diminuir os problemas dessa etapa na construção de um DWT, esse trabalho propõe uma ferramenta que realiza as atividades dessa fase. A ferramenta foi chamada de TETL, uma sigla do inglês *Trajectory's ETL*, ou seja, ETL de Trajetórias.

Kimball and Caserta (2004) afirmam que um sistema de ETL, adequadamente concebido, extrai os dados dos sistemas transacionais, enfatizando a qualidade dos dados e os padrões de consistência, validando-os de modo que fontes de dados distintas possam ser usadas juntas, e finalmente gerando os dados em um formato que as aplicações possam ler e usuários finais, tomar decisões.

De acordo com [KC04], a missão da equipe de ETL pode ser resumida em três funções básicas: (1) Entregar os dados de forma eficaz ao usuário final; (2) Adicionar valor aos dados, limpá-los e validá-los; (3) Proteger e documentar o mapeamento dos dados.

Esse capítulo objetiva apresentar aspectos relevantes sobre a ferramenta TETL, tanto no âmbito de seus requisitos como no que foi desenvolvido durante a execução do trabalho descrito nesse documento. Na seção 3.2, alguns detalhes principais sobre a ferramenta proposta são apresentadas, seus requisitos e principais objetivos, os conceitos adotados ou desenvolvidos, o formato de entrada dos dados, entre outros. Os aspectos sobre a implementação do protótipo durante o estudo realizado são encontrados na seção 3.3 desse capítulo, na qual estão informações sobre a definição de escopo, os componentes desenvolvidos, arquitetura, etc. A seção 3.4 destina-se a

apresentar algumas considerações finais sobre a ferramenta proposta, além de exibir um breve resumo sobre o que foi detalhado neste capítulo.

3.2 A ferramenta TETL

TETL é uma ferramenta de ETL para trajetórias de objetos móveis. O objetivo principal dessa ferramenta é executar cada uma das etapas do processo de ETL, que foi discutido na seção 2.4.2. Para alcançar este objetivo, a ferramenta deve criar uma base multidimensional a partir de dados extraídos de dispositivos capazes de receber informações do Sistema de Posicionamento Global, popularmente conhecido por GPS (do inglês *Global Positioning System*).

Durante a seção destinada a discutir o processo de ETL (seção 2.4.2), observou-se que ele estava diretamente relacionado ao modelo de dados multidimensional adotado no desenvolvimento do DW. Ficou evidente, após análise realizada na seção 2.4.1, que não existe um consenso acerca do modelo de dados sob o qual qualquer DWT deva ser desenvolvido. Esta falta de padronização existe tanto para as medidas como para as dimensões. Por esse motivo, para a construção da ferramenta TETL, um modelo de dados é proposto, considerando medidas previamente definidas, dimensões configuráveis e representação de dados oriundos de GPS. Esse modelo encontra-se descrito na seção 3.2.1.

Como existem inúmeros formatos de dados e por eles serem bastante distintos entre si, apenas um formato será considerado como entrada para TETL. Esse formato possui sua estrutura descrita na seção 3.2.2, que também inclui a justificativa da escolha desse formato de entrada de dados.

Após a apresentação do modelo de dados multidimensional adotado e do formato de entrada dos dados na ferramenta TETL, as etapas de ETL consideradas na ferramenta são descritas sistematicamente. Essa descrição é realizada na seção 3.2.2.

Portanto, esta seção está organizada como segue. O modelo multidimensional para dados de trajetória que é proposto neste documento é discutido na seção 3.2.1. A especificação e a justificativa sobre os aspectos relevantes acerca do formato de entrada dos dados na TETL são encontrados na seção 3.2.2. Já o processo de ETL adotado pela ferramenta está exposto na seção 3.2.3. Por fim, na seção 3.2.4, algumas conclusões sobre a ferramenta proposta e seus requisitos são expostos.

3.2.1 Modelo de dados multidimensional proposto

Devido ao fato de não haver uma convergência entre os pesquisadores da área abordada nesse documento sobre o modelo de dados multidimensional ideal para a criação de um DWT, TETL adota um modelo próprio. Esse modelo utiliza algumas características dos modelos apresentados na seção 2.4.1. A definição desse modelo pode ser visualizada na Figura 3.1.

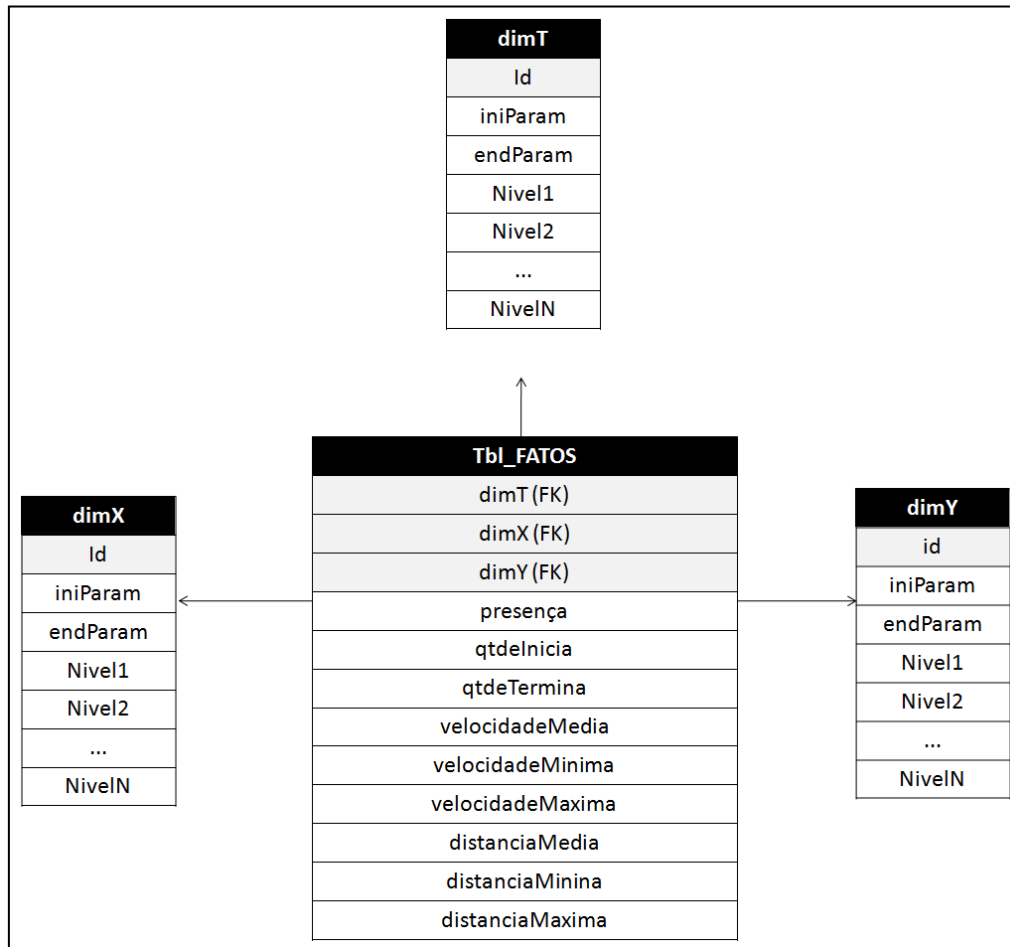


Figura 3.1: Modelo de dados adotado por TETL

O modelo dados multidimensional proposto é um modelo baseado nas coordenadas espaciais, por possuir duas dimensões, dimX e dimY, que correspondem às coordenadas longitudinais e latitudinais respectivamente. Ele possui também uma dimensão temporal, denominada de dimT.

A hierarquização das dimensões no modelo proposto é baseada em intervalos como exibido na Figura 2.5(b) e adotado no modelo de dados baseado em coordenadas espaciais. Como já dito em seções anteriores, a ferramenta TETL fornece ao usuário o

poder de configuração do intervalo que vai ser considerado mínimo e a quantidade de níveis hierárquicos de cada dimensão, permitindo, assim, a customização do resultado.

As medidas sugeridas para compor o modelo estão demonstradas na Tbl_FATOS da Figura 3.1, porém antes de especificar cada uma delas, o conceito de célula precisa ser reforçado. Uma célula é uma abstração do que é delimitado pelo cruzamento do intervalo temporal com os intervalos espaciais. Ou seja, uma célula é uma área espacial observada durante um determinado intervalo de tempo. Com esse conceito fundamentado, cada uma das medidas é descrita e especificada a seguir.

A medida **presença** representa a quantidade de trajetórias presentes na célula analisada. Assim como nos demais modelos analisados na seção 2.4.1, sua formulação obedece a especificação realizada por Braz *et al.* (2007).

As medidas **qtdeInicia** e **qtdeTermina** determinam a quantidade de trajetórias que iniciam ou terminam na célula em questão, respectivamente. Seus cálculos determinam se a primeira e a última observação da trajetória pertencem à célula em questão. A Figura 3.2 traz um exemplo de trajetória composta por seis observações. Os pontos a serem analisados para o cálculo das medidas qtdeInicia e qtdeTermina da trajetória representada na Figura 3.2 são os pontos P1 e P6, respectivamente.

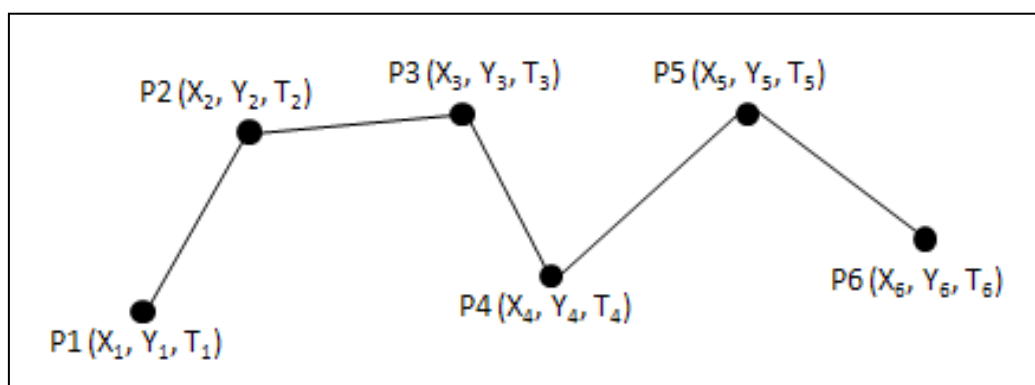


Figura 3.2: Exemplo de trajetória de um objeto móvel e suas observações

A distância percorrida por um objeto móvel em uma célula é a soma de todos os caminhos percorridos entre os pontos da trajetória que estão dentro da célula analisada. Ademais o tempo gasto por um objeto móvel é calculado pela diferença entre os momentos da primeira e da última observação da trajetória associada à célula em questão. Por exemplo, se a célula analisada incluísse os pontos P1, P2 e P3 da Figura 3.2, a distância percorrida por esse objeto na célula seria a soma das distâncias de P1 até

P2 e de P2 até P3, já o tempo gasto pelo objeto na mesma célula seria calculado pela diferença entre T_3 e T_1 .

As medidas relacionadas à distância são calculadas com base na explicação acima. As medidas ***distanciaMinima*** e ***distanciaMaxima*** são os valores mínimos e máximos dos valores das distâncias percorridas pelos objetos móveis presentes naquela célula. A medida ***distanciaMedia*** é calculada pela soma de todas as distâncias percorridas pelos objetos móveis presentes na célula, dividida pela quantidade de objetos móveis na célula, cujo valor é o da medida presença.

A fórmula para cálculo da velocidade é a mesma utilizada na Física: velocidade é igual à distância percorrida dividida pelo tempo gasto no percurso. Logo, as medidas ***velocidadeMinima*** e ***velocidadeMaxima*** determinam a menor e a maior velocidade de um objeto móvel presente na célula. Já a medida ***velocidadeMedia*** é calculada pela soma de todas as velocidades dos objetos móveis que passam por aquela célula, dividida pela medida presença.

Na seção seguinte, a análise comparativa entre as propostas de modelos de dados multidimensionais realizada na seção 2.4.1.1 é refeita para incluir o modelo de dados descrito e especificado nessa seção.

3.2.1.1 Análise comparativa entre os modelos existentes e o proposto

Na seção 2.4.1.3 foi realizada uma comparação entre os modelos de dados multidimensionais já propostos para DWT e discutidos nas seções anteriores. Após a apresentação do modelo proposto pelo presente documento, repetimos a análise feita naquela seção, incluindo o novo modelo para ser analisado segundo os mesmos critérios utilizados da primeira vez.

O Quadro 3.1 mostra que o modelo de dados multidimensional proposto combina as vantagens de cada um dos modelos já propostos e apresentados na seção 2.4.1, exceto a representação de propriedades dos objetos móveis envolvidos.

As informações dos objetos móveis não foram colocadas no modelo usado por TETL devido ao fato de ser uma informação muito particular da aplicação modelada. Um exemplo disso é uma aplicação na área de controle de tráfego aéreo que é bastante

diferente de uma aplicação de monitoramento de criminosos. Ambas as áreas possuem interesse em construir um DWT no seu domínio, porém suas modelagens são distintas.

Quadro 3.1: *Análise comparativa dos modelos multidimensionais existentes versus o proposto*

Crítérios	Coordenadas espaciais	Informações dos objetos móveis	Modelo usado por TETL
Representação de atributos de objetos móveis		✓	
Definição das medidas propostas		✓	✓
Dimensões configuráveis	✓	✓	✓
Representação de dados de GPS	✓		✓
Utilização em uma ferramenta de ETL para dados de GPS			✓

As medidas propostas pelo modelo de dados utilizado por TETL foram todas definidas formalmente. Isso favorece a adoção desse modelo por outros processos, pois facilita o entendimento e a implementação dos cálculos necessários durante as fases de transformação e carga.

Na Figura 3.1, nota-se que as dimensões não possuem uma quantidade definida de níveis da hierarquia. Assim como proposto nos demais modelos, no modelo proposto propõe que a hierarquização seja configurada pelo usuário, tornando o modelo multidimensional final customizado para atender as necessidades da aplicação.

Por fim, o modelo proposto é usado em uma ferramenta que provê suporte a dados extraídos de GPS, que é o principal sistema de posicionamento utilizado atualmente. Essa característica torna o modelo uma solução que pode ser adotada para a representação do posicionamento de diversos tipos de objetos móveis acoplados com receptores de GPS. Até o momento não se tem conhecimento sobre a existência de ferramentas de ETL que usem dados de GPS como entrada sem a necessidade de envolvimento do especialista no processo de ETL ou de programação para interpretação deste tipo de entrada de dados.

3.2.2 Fonte e formato dos dados de entrada

TETL utiliza como fonte de dados, arquivos de GPS. Esses arquivos possuem um formato específico, possuindo informações sobre o posicionamento espacial do objeto móvel em um determinado momento, além de outros dados, cuja importância depende da aplicação que faz uso deles.

Muitas interfaces recebem e trabalham os sinais recebidos nos seus receptores. Cada uma com suas particularidades e restrições. Uma delas é a NMEA (*National Marine Electronics Association*), que permite o desenvolvimento de sistemas distribuídos no contexto de computação móvel.

As mensagens NMEA são geradas constantemente por uma gama de receptores GPS. Essas mensagens podem ser tratadas por qualquer aplicação ou utilizadas *ad-hoc*. O uso dessas mensagens por outros computadores, em tempo real, permite o desenvolvimento de inúmeras aplicações em vários contextos, como no cenário urbano, no qual a aquisição de posicionamento geográfico é fundamental para várias atividades, como rastreamento de veículos ou pacotes.

A interface NMEA gera vários formatos de mensagens e, o mais simples deles é a RMC. Um exemplo de uma mensagem RMC pode ser vista na Figura 3.3 e a descrição da estrutura básica de uma mensagem desse tipo pode ser vista no Quadro 3.1.

```
$GPRMC,225446,A,4916.45,N,12311.12,W,1,054.7,191194,020.3,E*68
```

Figura 3.3: Exemplo de mensagem NMEA RMC (adaptação de [NMEA09]).

TETL usa mensagens NMEA RMC como formato de dados de entrada. Cada arquivo possui informações sobre o percurso de um objeto móvel. Portanto, para integrar trajetórias de vários objetos móveis, a ferramenta recebe mais de um arquivo, ou seja, é capaz de tratar dados de mais de um objeto e suas respectivas trajetórias.

A escolha de mensagens NMEA RMC como formato de entrada de dados na ferramenta TETL objetiva mostrar que, mesmo utilizando a mensagem mais simples, nove medidas podem ser extraídas. Portanto, para a geração de informação e conhecimento não é necessário que o formato de entrada dos dados possua dados complexos.

Quadro 3.2: Estrutura da mensagem NMEA RMC [NMEA09].

Parâmetro	Descrição	Valor na Figura 3.3
1°	Formato do dado (\$GPRMC ou \$RCM)	\$GPRCM
2°	Hora	225446 (22h 54min 46seg UTC)
3°	Estado dos dados (A = OK, V = Alerta)	A
4°	Latitude	4916.45 (49° 16' 45'')
5°	Direção Vertical (N = Norte, S = Sul)	N
6°	Longitude	12311.12 (123° 11' 12'')
7°	Direção Horizontal (E = Leste, W = Oeste)	W
8°	Velocidade no plano (na unidade de nós)	1 nó (1,852 Km/h)
9°	Direção angular momentânea	054.7
10°	Data	191194 (19/11/94 UTC)
11°	Variação magnética (graus)	020.3 (20,3°)
12°	Checksum (sempre começa com *)	*68

A ferramenta, além dos dados de GPS, recebe como entrada os parâmetros de configuração das dimensões e a unidade adotada para as medidas que concentram informações sobre velocidade. A unidade de medida utilizada para as medidas sobre distância percorrida é diretamente relacionada com a unidade escolhida para velocidade, ou seja, se a unidade para velocidade for Km/h, a distância será calculada em quilômetros.

Para cada dimensão, o usuário fornecerá dois parâmetros: (1) a quantidade de níveis hierárquicos desejados para essa dimensão e (2) o intervalo considerado como a menor granularidade da dimensão. Para as dimensões espaciais, dimX e dimY, o intervalo será dado em minutos latitudinais ou longitudinais dependendo da dimensão em questão. Para a dimensão temporal, o intervalo será em minutos.

3.2.3 Etapas do processo de ETL da ferramenta TETL

Uma breve sistematização das etapas de ETL consideradas na ferramenta TETL pode ser observada na Figura 3.4. Nela é possível entender a execução de cada uma das etapas do processo de ETL proposto.

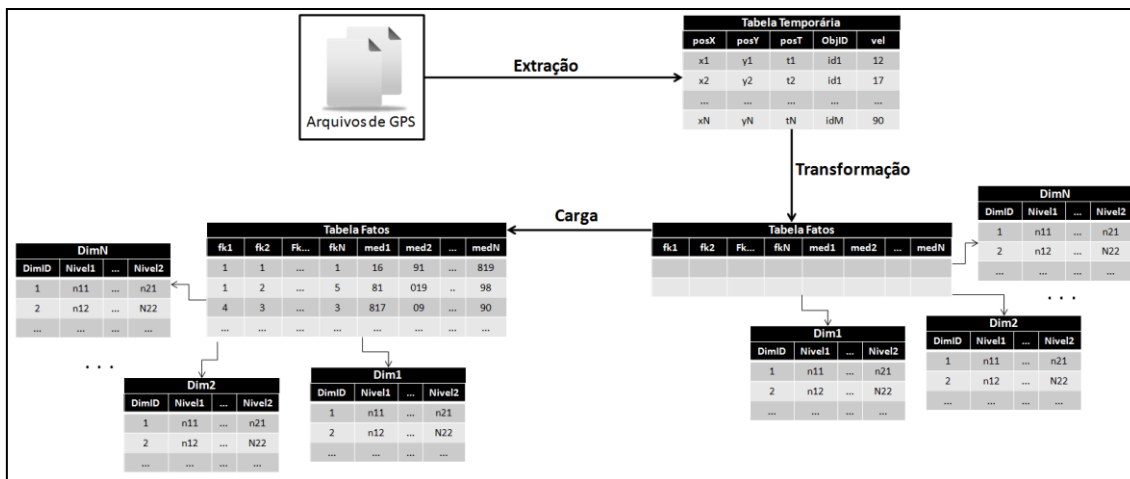


Figura 3.4: Etapas do processo de ETL na ferramenta TETL

A entrada da ferramenta é composta de vários arquivos de GPS, os quais possuem informações sobre o posicionamento espacial dos objetos no tempo. O formato escolhido foi apresentado na seção 3.2.2.

Esses dados passam pelo processo de extração, no qual os dados são armazenados em uma tabela temporária, alinhando apenas as informações necessárias, como: posição latitudinal, longitudinal e temporal, além da velocidade do objeto nesse instante. Como são utilizados mais de um arquivo como fonte de dados, e cada arquivo corresponde às medidas de um único objeto móvel, cria-se um identificador do objeto móvel para que não se misturem as trajetórias de objetos distintos.

Durante a etapa de transformação, algumas conversões são realizadas e a hierarquização das dimensões já começa a ser montada. A decisão do menor nível hierárquico, da quantidade de hierarquias e das unidades de medida a serem usadas em cada uma das dimensões (tanto as espaciais como a temporal) é de responsabilidade do usuário. Para isso, o processo deve permitir que o usuário realize sua própria configuração, para que a modelagem multidimensional obtida ao final do processo atenda às suas necessidades.

Após essas definições terem sido realizadas durante a etapa de transformação, é o momento da realização da próxima etapa, a de carga dos dados na base multidimensional, momento no qual os dados são finalmente inseridos na base modelada multidimensionalmente.

Existem três maneiras de realizar a carga dos dados durante o processo de ETL: (1) Carga Total: todos os dados são inseridos de uma única vez; (2) Carga de

atualização: os dados são truncados e completamente recarregados; e (3) Carga incremental: os dados que já estão na base devem ser mantidos sempre que novos dados são adicionados. TETL, por questões de escopo, faz uso da primeira forma de inserção de dados no DW: carga total. Assim, sempre que a ferramenta for executada com dados novos, uma base diferente será gerada, a qual não contará com os dados já armazenados anteriormente.

3.3 Protótipo da ferramenta TETL

A ferramenta TETL foi especificada na seção 3.2. Porém, para ilustrar a aplicação das propostas realizadas, precisa-se que um protótipo da ferramenta seja desenvolvido, implementando os requisitos dispostos durante o seu detalhamento. O objetivo principal desta seção é apresentar os aspectos envolvidos no protótipo criado.

A definição da arquitetura de software sobre a qual o processo de ETL agirá é um ponto fundamental para que o resultado deste processo seja satisfatório. Outro ponto importante na especificação da arquitetura é definir se o processo de ETL será implementado com o apoio de uma ferramenta preexistente ou a partir da implementação manual de todo o processo [KC04].

O uso de uma ferramenta já existentes no mercado e na academia possui a vantagem de suportar vários tipos de dados. Porém, nenhuma delas fornece suporte a dados de GPS. Para que este tipo de dados fosse suportado pelo processo, uma extensão de alguma ferramenta de código aberto pode ser realizada. A implementação dessa extensão exigiria algumas alterações no código da ferramenta, as quais demandariam tempo para entendimento do código e tempo para desenvolvimento.

A decisão de desenvolver a ferramenta TETL a partir de código próprio se deve ao fato de que os requisitos do processo de ETL podem ser alterados e essas alterações exigem manutenção no código. Para a realização da manutenção do código, o envolvimento de um especialista no processo seria obrigatório, em ambas as possibilidades, tanto a de estender uma ferramenta de código aberto como a de criar uma de código próprio. Porém, o custo de manter código próprio é inferior ao de atualizar uma ferramenta de código aberto.

O protótipo desenvolvido cobre cada uma das etapas descritas para o processo adotado. Portanto, o protótipo é capaz de receber os arquivos de GPS e os parâmetros

escolhidos pelo usuário, como especificados na seção 3.2.2, realizar cada uma das etapas do processo descrito na seção 3.2.3 e entregar ao usuário uma base multidimensional de dados modelada segundo o modelo proposto e descrito na seção 3.2.1.

Para descrever o protótipo desenvolvido, esta seção está organizada como segue. A seção 3.3.1 apresenta a arquitetura sobre a qual o protótipo da ferramenta foi desenvolvido. Na seção 3.3.2 estão descritos os aspectos acerca das tecnologias utilizadas no desenvolvimento da ferramenta.

3.3.1 Arquitetura

Kimball and Caserta (2004) alertam que para a ferramenta de ETL ser realmente eficiente, sua arquitetura deve ser bem definida e modularizada. Isso se deve ao fato de que o processo realizado pela ferramenta pode sofrer mudanças, e as alterações necessárias para a implementação dessas mudanças precisam ser de fácil identificação e implementação. Ou seja, a arquitetura do sistema deve estar de acordo com a proposta inicial, mas precisa ser flexível o bastante para que possíveis alterações possam ser realizadas com o menor esforço possível.

A Figura 3.5 exibe a arquitetura do protótipo da ferramenta TETL. As responsabilidades de cada componente são explicadas a seguir. O componente **DatabaseCreator** é um componente responsável por criar o esquema inicial da base de dados que vai ser utilizada como destino dos dados. Ele cria a tabela temporária que receberá os dados de cada informação analisada e as tabelas do modelo multidimensional, tanto as de dimensões como a de fatos. Esse componente também é o responsável por criar as tabelas de dimensões de acordo com a configuração desejada pelo usuário para a hierarquização de cada dimensão.

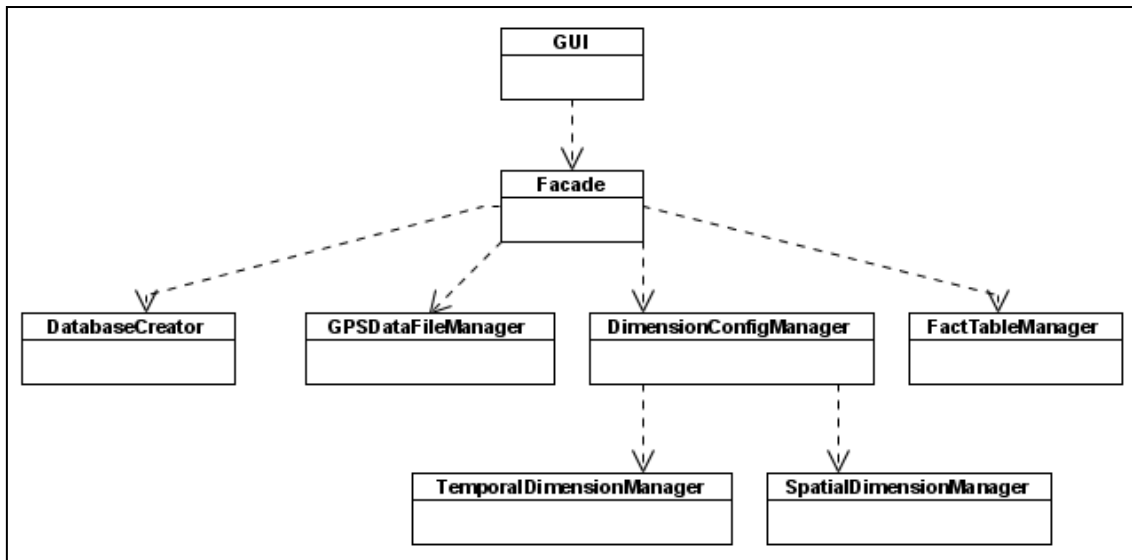


Figura 3.5: Arquitetura do protótipo da ferramenta TETL

O componente ***GPSDataFileManager*** é o responsável pela leitura dos arquivos de dados de GPS que são fornecidos como entrada para a execução da ferramenta. Já o componente ***DimensionConfigManager*** é o responsável por inserir as instâncias de cada dimensão, respeitando os parâmetros escolhidos pelo usuário. Ele utiliza dois componentes auxiliares para a execução de suas atividades: ***TemporalDimensionManager*** e ***SpatialDimensionManager***. Esses componentes são os responsáveis por calcular os intervalos de tempo e espaço, respectivamente, e inserir efetivamente os dados na base.

O último componente, o ***FactTableManager***, é o responsável por calcular as medidas de acordo com os valores das dimensões e inserir os dados na tabela de fatos.

Todos os componentes citados, exceto o ***GPSDataFileManager***, são abstratos e suas implementações são diretamente dependentes do SGBD escolhido para armazenamento de dados, já que realizam operações com a base de dados.

3.3.2 Tecnologias envolvidas

A tecnologia escolhida para desenvolver o protótipo da ferramenta foi Java. Essa escolha se deve ao fato de Java ser uma linguagem muito bem documentada e utilizada em larga escala tanto no mercado como na academia. Essas características favorecem a resolução de problemas que possam acontecer durante a etapa de projeto da ferramenta.

Como pode ser observado, o resultado obtido ao final do processo da ferramenta é uma base de dados multidimensional customizada de acordo com os parâmetros

escolhidos pelo usuário da aplicação. Essa base de dados precisa ser criada de acordo com as regras de SGBD já existente no mercado, tais como: Oracle, MySQL e PostgreSQL.

Para a implementação do protótipo, o SGBD escolhido foi o PostgreSQL. Essa escolha foi feita devido ao PostgreSQL ser um SGBD gratuito, de fácil instalação e extensível. Porém, como foi analisado na seção anterior, o sistema possui uma arquitetura modular para que possa ser mais facilmente estendido para outros SGBD e, nas próximas versões, prover ao usuário, mais essa possibilidade de configuração.

3.4 Conclusões

Devido ao fato de o desenvolvimento de DWT ser uma área de pesquisa que precisa de muito investimento e estudo, algumas etapas do processo de criação ainda precisam ser especificadas. Uma delas é a etapa de ETL. Para resolver esse problema, este trabalho propõe a ferramenta TETL, a qual foi detalhada neste capítulo. TETL é uma ferramenta que objetiva extrair informações agregadas de dados de GPS, funcionalidade não fornecida por nenhuma outra ferramenta de ETL até o momento.

Para a especificação de uma ferramenta de ETL é necessário que se defina o processo e o modelo de dados multidimensional adotados. O processo de ETL escolhido para ser executado na ferramenta TETL foi apresentado e detalhado na seção 3.2.3. Já o modelo de dados adotado para a representação multidimensional dos dados de trajetórias foi descrito na seção 3.2.1. Este modelo é uma nova proposta, que considera medidas previamente definidas, permite configuração das dimensões e representa dados oriundos de GPS.

Após a especificação da ferramenta, realizada na seção 3.2, o protótipo desenvolvido teve seus detalhes exibidos na seção 3.3. Este protótipo será utilizado para o estudo de caso desenvolvido e detalhado no próximo capítulo.

Capítulo 4

Estudo de caso utilizando TETL

4.1 Introdução

Este capítulo sintetiza as idéias expostas no capítulo anterior exibindo uma discussão sobre um estudo de caso, no qual foi feito uso do modelo de dados para DWT e do processo de ETL propostos neste documento. Ainda são ilustradas as funcionalidades do protótipo da ferramenta TETL. São utilizadas as propostas feitas e demonstração da ferramenta. Portanto este capítulo tem como objetivo a apresentação de uma aplicação de ETL por meio da ferramenta TETL.

A demonstração da ferramenta TETL será realizada sobre uma aplicação na área de rastreamento de uma frota de veículos. A descrição do estudo de caso escolhido, incluindo detalhes dos arquivos de entrada fornecidos à ferramenta, está disposta na seção 4.2. Na seção 4.3, o comportamento da ferramenta durante cada uma das etapas do processo de ETL adotado pela ferramenta é apresentado. Por fim, algumas considerações finais sobre a validação do trabalho realizado são dadas na seção 4.4.

4.2 Estudo de caso analisado

O caso de estudo escolhido é no domínio de rastreamento de veículos. A Figura 4.1 exibe o cenário envolvido no caso analisado. A frota de automóveis rastreados realiza seus percursos utilizando um receptor de sinal de GPS que possuam a funcionalidade de emitir mensagens NMEA RCM, como as especificadas na seção 3.2.1. As mensagens emitidas por cada veículo são armazenadas em arquivos separados. Esses arquivos são utilizados como entrada para a ferramenta.

Como foi especificado no capítulo anterior, TETL recebe, além dos arquivos de GPS, parâmetros de configuração. Esta configuração é responsável por determinar a unidade na qual as medidas de agregação serão calculadas, além do intervalo mínimo e da quantidade de níveis na hierarquia de cada dimensão. Esses parâmetros são de inteira

responsabilidade do usuário final da aplicação e é partir deles que o resultado final será customizado.

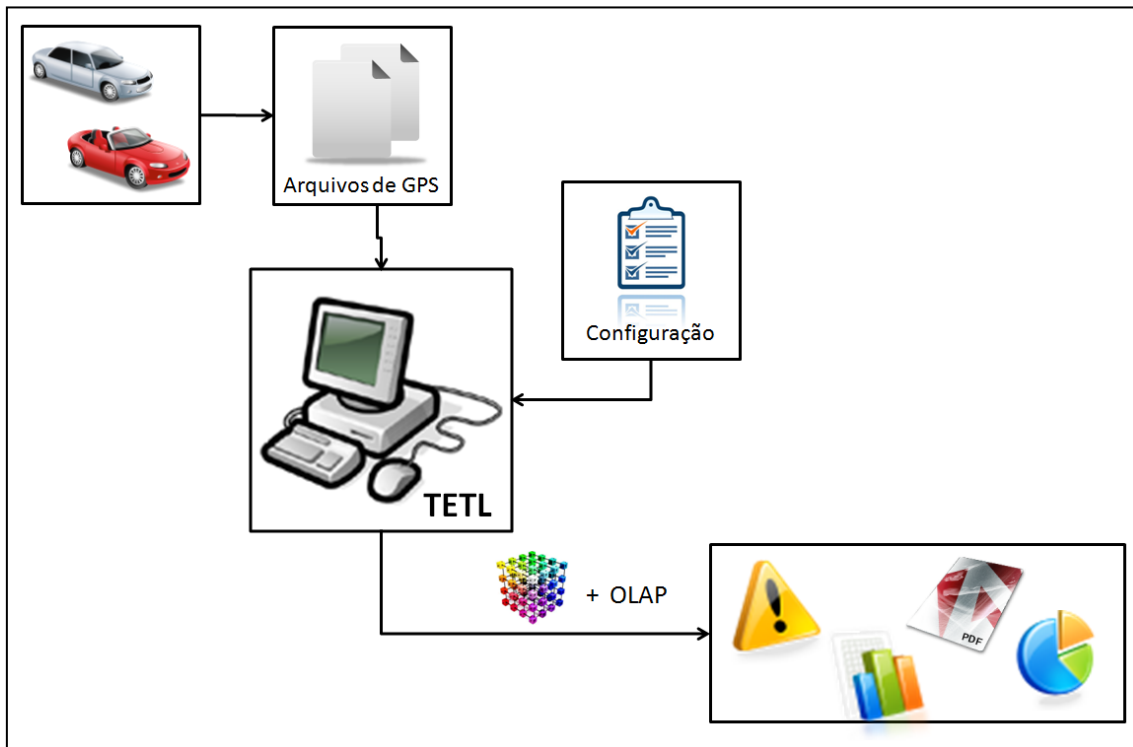


Figura 4.1: Cenário do estudo de caso utilizando TETL

A partir dessas entradas (o conjunto de arquivos de GPS e os parâmetros de configuração), a ferramenta pode ser executada. Seu funcionamento poderá ser acompanhado na próxima seção, a 4.3. O resultado final será uma base de dados modelada multidimensionalmente e povoada com as informações extraídas dos arquivos de GPS e customizada a partir da configuração desejada.

O resultado poderá ser utilizado como fonte para a geração de cubos e realização de consultas OLAP. Ao final dessa etapa, relatórios, gráficos, índices e alertas poderão ser encontrados de forma mais fácil. A ferramenta TETL não possui suporte a essas duas últimas etapas (geração de cubo e consultas OLAP), pois foge ao escopo do sistema.

Os dados de GPS utilizados como entrada para esse estudo de caso são reais e foram extraídos de uma frota de veículos da Alemanha. São dados sobre 102 veículos observados entre as datas 5 e 10 de novembro de 2004. A Figura 4.2 traz um trecho de um dos arquivos utilizados como entrada na execução desse estudo de caso. Todos os arquivos estão formatados de maneira semelhante: mensagens NMEA RMC escritas

uma após a outra. Cada arquivo possui informações sobre a trajetória de um único objeto móvel e cada linha representa uma observação realizada.

```
$GPRMC,170257.824,A,5154.0414,N,00738.1837,E,5.75,77.74,051104,,*3B
$GPRMC,170258.824,A,5154.0424,N,00738.1864,E,7.09,57.38,051104,,*32
$GPRMC,170259.824,A,5154.0438,N,00738.1893,E,8.29,49.49,051104,,*32
$GPRMC,170300.824,A,5154.0453,N,00738.1924,E,8.84,49.94,051104,,*38
$GPRMC,170301.824,A,5154.0471,N,00738.1955,E,9.42,47.09,051104,,*3E
$GPRMC,170302.824,A,5154.0484,N,00738.1995,E,10.12,60.90,051104,,*03
$GPRMC,170303.824,A,5154.0496,N,00738.2039,E,10.62,66.19,051104,,*0D
$GPRMC,170304.824,A,5154.0506,N,00738.2085,E,10.79,69.57,051104,,*0A
$GPRMC,170305.823,A,5154.0518,N,00738.2130,E,10.99,66.21,051104,,*0C
$GPRMC,170306.823,A,5154.0531,N,00738.2175,E,11.04,64.53,051104,,*07
$GPRMC,170307.823,A,5154.0545,N,00738.2220,E,11.23,63.88,051104,,*02
$GPRMC,170308.823,A,5154.0559,N,00738.2266,E,11.31,63.58,051104,,*0C
$GPRMC,170309.823,A,5154.0567,N,00738.2299,E,7.47,69.71,051104,,*37
$GPRMC,170310.823,A,5154.0582,N,00738.2342,E,10.58,63.40,051104,,*03
```

Figura 4.2: Trecho de arquivo de GPS utilizado como entrada no estudo de caso

4.3 TETL em funcionamento

Esta seção objetiva demonstrar da ferramenta TETL em funcionamento. Na seção 4.3.1 a interface do protótipo e o detalhamento dos parâmetros escolhidos para a execução do estudo de caso são apresentados. Na seção 4.3.2, os resultados obtidos após a execução de cada uma das etapas do processo adotado pela ferramenta TETL são exibidos.

4.3.1 Interface com o usuário

O objetivo do protótipo da ferramenta TETL é desenvolver todas as etapas do processo adotado e respeitar todos os requisitos de software do sistema. Optou-se pela implementação de uma aplicação *desktop* que consiga obter do usuário todas as informações que precisam ser passadas por ele. Essas informações são o diretório de armazenamento dos arquivos de GPS obtidos de cada um dos veículos rastreados pela frota e os parâmetros de configuração escolhidos pelo usuário.

A Figura 4.3 ilustra a interface com o usuário do protótipo da ferramenta TETL. Nela é possível identificar os campos nos quais o usuário fornece as informações necessárias para a execução da ferramenta.

Figura 4.3: Interface com o usuário do protótipo da ferramenta TETL

Os parâmetros escolhidos para a execução desse estudo de caso também podem ser visualizados na Figura 4.3. Os intervalos das dimensões espaciais foram escolhidos para 1' (um minuto), pois como a área percorrida foi a Alemanha, o intervalo espacial foi pequeno. E o intervalo de 5 horas como o intervalo mínimo da dimensão temporal se deve ao fato de o tempo observado ter sido de apenas 6 dias. Já as quantidades de níveis hierárquicos de cada dimensão foram escolhidas aleatoriamente.

4.3.2 Execução do processo

Essa seção objetiva demonstrar a execução de cada uma das etapas do processo descrito na seção 3.2.3, a qual descreveu o processo utilizado na ferramenta TETL. A primeira etapa do processo é a inserção dos dados extraídos dos arquivos de GPS em uma tabela temporária. Um trecho dos dados dessa tabela pode ser visualizado na Figura 4.4. Foram geradas 62.890 entradas para essa tabela, ou seja, durante o período analisado foram realizadas exatamente esse número de observações sobre os 102 veículos da frota

	objid integer	latitude double precis	longitude double precis	moment date	velocity double precision
1	2	51.95	7.63	2004-11-08	0.0831999969482422
2	2	51.95	7.63	2004-11-08	0.0779999971389771
3	3	51.97	7.63	2004-11-06	0.0415999984741211
4	3	51.97	7.63	2004-11-06	0.036399998664856
5	3	51.97	7.63	2004-11-06	0.0519999980926514
6	3	51.97	7.63	2004-11-06	0.0311999988555908
7	3	51.97	7.63	2004-11-06	0.0259999990463257
8	3	51.97	7.63	2004-11-06	0.036399998664856
9	3	51.97	7.63	2004-11-06	0.0311999988555908

Figura 4.4: Trecho da tabela temporária

A próxima etapa é responsável por povoar as dimensões, calculando os intervalos de granularidade mínima. A Figura 4.5 exibe trechos de cada uma das tabelas de dimensões. A Figura 4.5 (a) mostra os dados da dimensão dimX, a qual foi povoada com 35 instâncias. Já a Figura 4.5 (b) exibe parte dos dados da dimY, que foi preenchida por 162 instâncias. Por fim, a Figura 4.5 (c) expõe parte da tabela representativa da dimensão dimT, que foi povoada por 32 instâncias.

(a)	id [PK] integer	iniparam double precis	endparam double precis	level1 character(50)	level2 character(50)	level3 character(50)
	11	2.74	3.26	2.74 -> 3.26		
	12	3.26	3.46	3.26 -> 3.46		
	13	3.46	3.79	3.46 -> 3.79		
	14	3.79	4.34	3.79 -> 4.34		
	15	4.34	4.59	4.34 -> 4.59		
	16	4.59	5.1	4.59 -> 5.1		

(b)	id [PK] integer	iniparam double precis	endparam double precis	level1 character(50)	level2 character(50)	level3 character(50)
	72	22.84	23.43	22.84 -> 23.43		
	73	23.43	23.74	23.43 -> 23.74		
	74	23.74	24.26	23.74 -> 24.26		
	75	24.26	24.46	24.26 -> 24.46		
	76	24.46	24.79	24.46 -> 24.79		
	77	24.79	25.34	24.79 -> 25.34		

(c)	id [PK] integer	iniparam date	endparam date	level1 character(50)	level2 character(50)	level3 character(50)
	9	2004-11-06	2004-11-06	06/11/2004 02:11:36 -> 06/11/2004 07:11:36		
	10	2004-11-06	2004-11-06	06/11/2004 07:11:37 -> 06/11/2004 12:11:37		
	11	2004-11-06	2004-11-06	06/11/2004 12:11:38 -> 06/11/2004 05:11:38		
	12	2004-11-06	2004-11-06	06/11/2004 05:11:39 -> 06/11/2004 10:11:39		
	13	2004-11-06	2004-11-07	06/11/2004 10:11:40 -> 07/11/2004 03:11:40		
	14	2004-11-06	2004-11-07	07/11/2004 03:11:41 -> 07/11/2004 08:11:41		
	15	2004-11-06	2004-11-07	07/11/2004 08:11:42 -> 07/11/2004 01:11:42		

Figura 4.5: Trecho das tabelas de dimensão

A partir das dimensões já determinadas, a última etapa do processo é o preenchimento da tabela de fatos com as medidas de agregação adotadas pelo modelo proposto. A Figura 4.6 exibe parte das 81 instâncias dos fatos mensurados pela aplicação. A baixa quantidade de entradas na tabela de fatos se deve ao fato de que células vazias, ou seja, sem nenhuma observação, são desconsideradas pela ferramenta.

dimxfk [PK] int	dimyfk [PK] int	dimtfk [PK] int	maxspeed double precis	minspeed double	avgspeed double	maxdist double	mindist double	avgdist double	presence integer	startsin integer	endsin integer
26	157	26	2203.096481920	0	1101.54	852598.330	0	0	2	1	2
26	158	6	4.940656458410	0	0	0	0	0	6	6	6
26	158	10	11904.04050950	0	2414.58	2069052.80	0	0	6	5	3
26	158	11	134.6815561100	0	15.7747	3059.7200	0	0	11	10	11
26	158	14	580.518805785	580.5188	580.518	137002.430	137002.0	0	1	1	0
26	158	19	3370.892150340	0	1685.44	684291.100	0	0	2	0	1
26	158	20	4985.929683630	0	2577.02	1625482.50	0	0	3	3	2
26	158	22	2249.29129958	2249.291	2249.29	1363070.50	1363070.0	0	1	0	0

Figura 4.6: Trecho da tabela de fatos

4.4 Conclusões

Após a apresentação deste capítulo, conclui-se que os objetivos e requisitos propostos para a ferramenta são respeitados. TETL é capaz de gerar dados multidimensionais a partir de dados de GPS, os quais não eram utilizados anteriormente por nenhuma outra aplicação de ETL.

Com a utilização do protótipo da ferramenta TETL, o qual foi validado por meio de um estudo de caso neste capítulo, a etapa mais crítica da criação de um *Data Warehouse* de trajetórias está finalizada e implementada. Algumas melhorias podem ser feitas na ferramenta proposta. Essas melhorias estão descritas na seção 5.3

Capítulo 5

Considerações Finais

5.1. Introdução

A apresentação dos principais conceitos da área de trajetórias de objetos móveis foi realizada no Capítulo 2. Já a especificação e os detalhes de implementação e funcionamento da ferramenta TETL, que se propõe a realizar as etapas do processo de ETL para dados de GPS podem ser encontrados nos capítulos 3 e 4, respectivamente.

Este capítulo tem como objetivo exibir um resumo sobre o trabalho realizado e listar as considerações finais sobre os principais aspectos discutidos neste documento. Na seção 5.2, são listadas as principais contribuições deixadas pelo trabalho apresentado nesse documento. Por fim, a seção 5.3 traz algumas indicações de trabalhos futuros que podem ser realizadas para o aprimoramento do estudo feito.

5.2. Principais contribuições

Como principais contribuições do trabalho aqui documentado, primeiramente, pode-se citar a especificação e implementação de uma ferramenta automática para a realização do processo de ETL, que é considerado a etapa mais crítica na criação de qualquer *Data Warehouse*, inclusive para um DW de trajetórias. A partir da utilização da ferramenta, dados de GPS que não geravam informação agregada podem agora ser utilizados como fonte de Sistemas de Suporte à Decisão.

A ferramenta criada não só possibilita a utilização dos dados de GPS modelados multidimensionalmente, como permite que o resultado obtido tenha mais utilidade para o domínio de aplicação escolhido. Isto é conseguido por permitir que a configuração do modelo de dados multidimensional seja realizada pelo usuário. Esta configuração torna o resultado mais flexível e customizado para cada usuário, portanto, atendendo melhor às suas necessidades.

Houve ainda a especificação de um processo de ETL para dados de GPS, utilizando esses dados que antes não eram utilizados por nenhuma ferramenta

conhecida. Esse processo foi o adotado no desenvolvimento da ferramenta TETL. Além da criação dessa ferramenta, uma análise detalhada de vários aspectos envolvidos no contexto do trabalho foi apresentada neste documento. Foram realizadas, ainda, uma categorização dos modelos de dados multidimensionais encontrados na literatura e uma comparação entre os modelos de cada uma das categorias criadas foi realizada. A partir dessa comparação, um modelo de dados para DWT que considera medidas previamente definidas, dimensões configuráveis e representação de dados oriundos de GPS foi proposto na seção 3.2.1.

5.3. Trabalhos futuros

Após a análise apresentada neste documento, é possível identificar melhorias que não foram tratadas, mas que agregariam bastante valor à ferramenta TETL se fossem implementadas. Entre elas, podem ser citadas:

1. O processo de ETL realizado pela ferramenta pode ser realizado por meio de uma carga incremental, o que aumentaria a aplicabilidade da solução proposta em várias áreas;
2. A ferramenta pode ser estendida para aceitar mais de um formato de representação de localização como entrada, não restringindo a entrada de arquivos no formato NMEA RCM;
3. Aplicar ao processo de ETL, os conceitos de reconstituição de trajetórias especificados por Braz *et al.* (2007) aumentaria a precisão dos cálculos das medidas de agregação calculadas a partir da distância percorrida pelo objeto. A Figura 5.1 (a) apresenta os pontos de uma trajetória ligados por retas. Já na Figura 5.1 (b), alguns pontos são adicionados a trajetória a partir da proposta de reconstituição feita por estes autores;

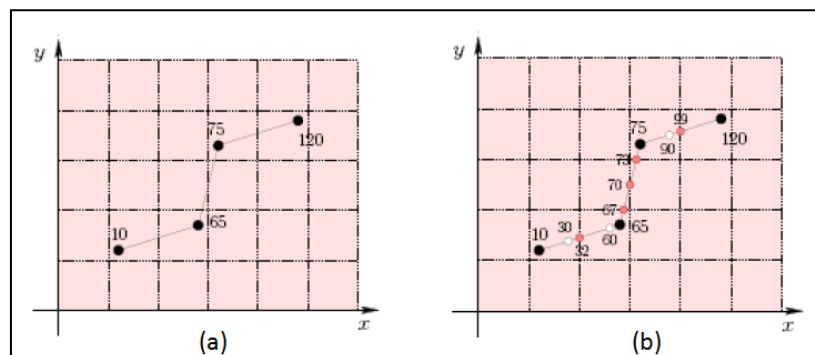


Figura 5.1: Proposta de reconstituição de trajetórias ([BRAZ07])

4. Acrescentar suporte a semântica de trajetórias, aumentando assim a utilidade da ferramenta em diversas áreas;
5. Permitir que a configuração do SGBD utilizado como base da ferramenta seja feita pelo usuário;
6. Investigar e incrementar a ferramenta com a melhor maneira de realizar as agregações das dimensões;
7. Há ainda a possibilidade de correção de *outliers* de trajetórias. O conceito de *outlier* representa um dado grosseiramente diferente ou inconsistente com os demais dados de um conjunto [HK06]. A Figura 5.2 mostra um *outlier* na trajetória TR3 no conjunto das trajetórias analisadas.

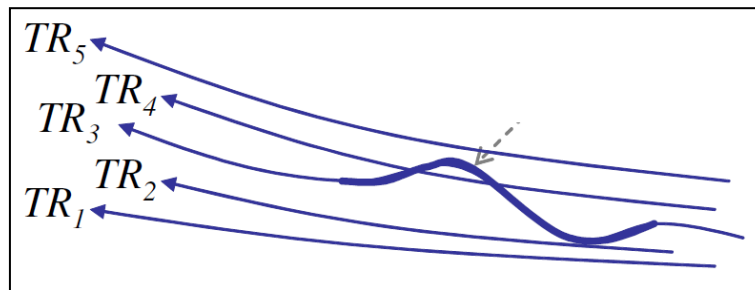


Figura 5.2: Exemplo de outlier em uma trajetória (adaptação de [LHL08])

Referências

- [ABKMMV07] Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., and Vaisman, A. (2007). **A model for enriching trajectories with semantic geographical information.** In ACM-GIS, New York, NY, USA. ACM Press, 162-169.
- [BDH08] Baltzer, O., Dehne, F., Hambrusch, S. (2008) **OLAP for Trajectories.** In 19th International Conference on Database and Expert Systems Applications, DEXA 2008, 340-347.
- [BOORRS06] Braz F., Orlando S., Orsini R., Raffaetà A., Roncato A., Silvestri C. (2006) **Towards a Trajectory Data Warehouse.** In Technical Report CS-2006-4, Università Ca' Foscari di Venezia.
- [BOORRS07] Braz, F., Orlando, S., Orsini, R., Raffaetà, A., Roncato, A., and Silvestri, C. (2007). **Approximate Aggregations in Trajectory Data Warehouses.** In STDM Workshop. 536-545.
- [BOWA08] Bogorny, V. and Wachowicz, M. (2008). **A Framework for Context-Aware Trajectory Data Mining.** Springer.
- [BRAZ07] Fernando J. Braz. (2007) **Trajectory Data Warehouses: Proposal of Design and Application to Exploit Data.** In 9th Brazilian Symposium on Geoinformatics, Campos do Jordão, Brazil, November 25-28, INPE, 61-72.
- [CHDAY97] Chaudhuri S., Dayal U. (1997) **An overview of data warehousing and OLAP technology.** In ACM SIGMOD Record, v.26 n.1, 65-74.
- [ELNA03] Elmasri, R. and Navathe, S. B. (2003). **Fundamentals of Database Systems.** Fourth Edition. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

- [HK06] Han, J., Kamber, M. (2006). **Data Mining: Concepts and Techniques**. 2nd ed. Morgan Kaufmann.
- [INM05] W. H. Inmon. (2005) **Building the Data Warehouse**. 4th Edition. John Wiley & Sons, Inc.
- [KAI] Kai India. Disponível em: <http://kaiindia.co.in/index_sub3.html>. Acesso em: 19 out. 2009.
- [KC04] Kimball, R., Caserta, J. (2004) **The Data Warehouse Etl Toolkit: Practical Techniques for Extracting**. Cleanin. John Wiley & Sons, Inc.
- [LHL08] Jae-Gil Lee, Jiawei Han, Xiaolei Li. (2008) **Trajectory Outlier Detection: A Partition-and-Detect Framework**. In ICDE 2008. IEEE 24th International
- [LORRS09] Leonard, L., Orlando, S., Raffaetà, A., Roncato, A. Silvestri, C. (2009) **Frequent Spatio-Temporal Patterns in Trajectory Data Warehouses**. In SAC'09 March 8-12, 2009, Honolulu, Hawaii, U.S.A.
- [MARK08] Marketos, G., Frentzos, E., Ntoutsis, I., Pelekis, N., Raffaetà, A., and Theodoridis, T. (2008) **Building Real-World Trajectory Warehouses**. In *MobiDE*, 8-15.
- [NIS03] Nissen, G. (2003) **Is Hand-Coded ETL the Way to Go?** In Intelligent Enterprise magazine.
- [NMEA09] **NMEA Data**. Disponível em: <<http://www.gpsinformation.org/dale/nmea.htm>>. Acesso em: 16 nov. 2009.
- [OORR07] Orlando, S., Orsini, R. Raffaetà, A., Roncato, A. (2007a) **Trajectory Data Warehouse: Design and Implementation Issues**. In Journal of Computing Science and Engineering, Vol. 1, No. 2, December 2007, 211-232.

- [OORRS07] Orlando, S., Orsini, R., Raffaetà, A., Roncato, A., and Silvestri, C. (2007b). **Spatio-temporal aggregations in trajectory data warehouses.** In Song, I. Y., Eder, J., and Nguyen, T. M., editors, DaWaK, volume 4654 of Lecture Notes in Computer Science, 66–77. Springer.
- [PRDVMFNT08] Pelekis, N., Raffaetà, A., Damiani, M. L., Vangenot, C., Marketos, G., Frentzos, E., Ntoutsis, I., and Theodoridis, Y. (2008). **Towards trajectory data warehouses.** In Giannotti, F. and Pedreschi, D., editors, Mobility, Data Mining and Privacy, 189–211. Springer.
- [RK04] M. Ross, R. Kimball. (2004) **Surrounding The ETL Requirements.** In Intelligent Enterprise magazine.
- [SPAC08] Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., and Vangenot, C. (2008). **A conceptual view on trajectories.** Data and Knowledge Engineering, 65(1):126–146.
- [WOL98] Wolson, O., Xu, B., Chamberlain, S., Jiang, L. (1998). **Moving Objects Databases: Issues and Solutions.** In Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference.

In ACM-GIS'07, November 7-9, 2007, Seattle, WA.

Anexos

Anexo A – Imagens dos modelos de dados analisados

Por respeito à formatação do documento, as imagens dos modelos de dados analisados na seção 2.4.1 foram exibidas em tamanho reduzido. Essa redução prejudica a visualização de detalhes de cada modelo. Com o objetivo de resolver esse problema, este anexo é colocado junto ao documento.

A Figura A.1 ilustra o modelo proposto por Orlando *et al.* (2007) e a Figura A.2 traz o modelo proposto por Braz (2007). Ambos propõem um modelo baseado em coordenadas espaciais. Já na Figura A.3 é apresentada o modelo baseado nas informações dos objetos móveis proposto por Marketos (2008).

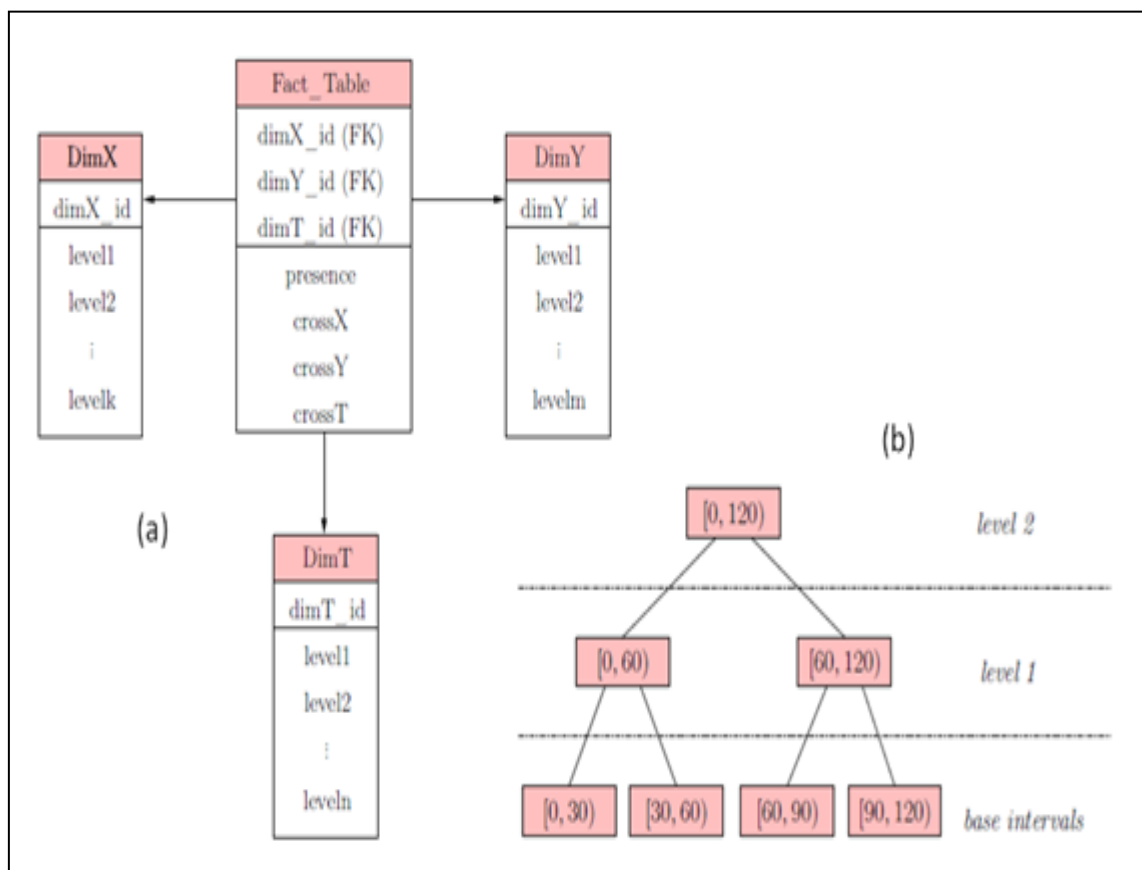


Figura A.1: Modelagem baseada em coordenadas espaciais (adaptação de [OORR07])

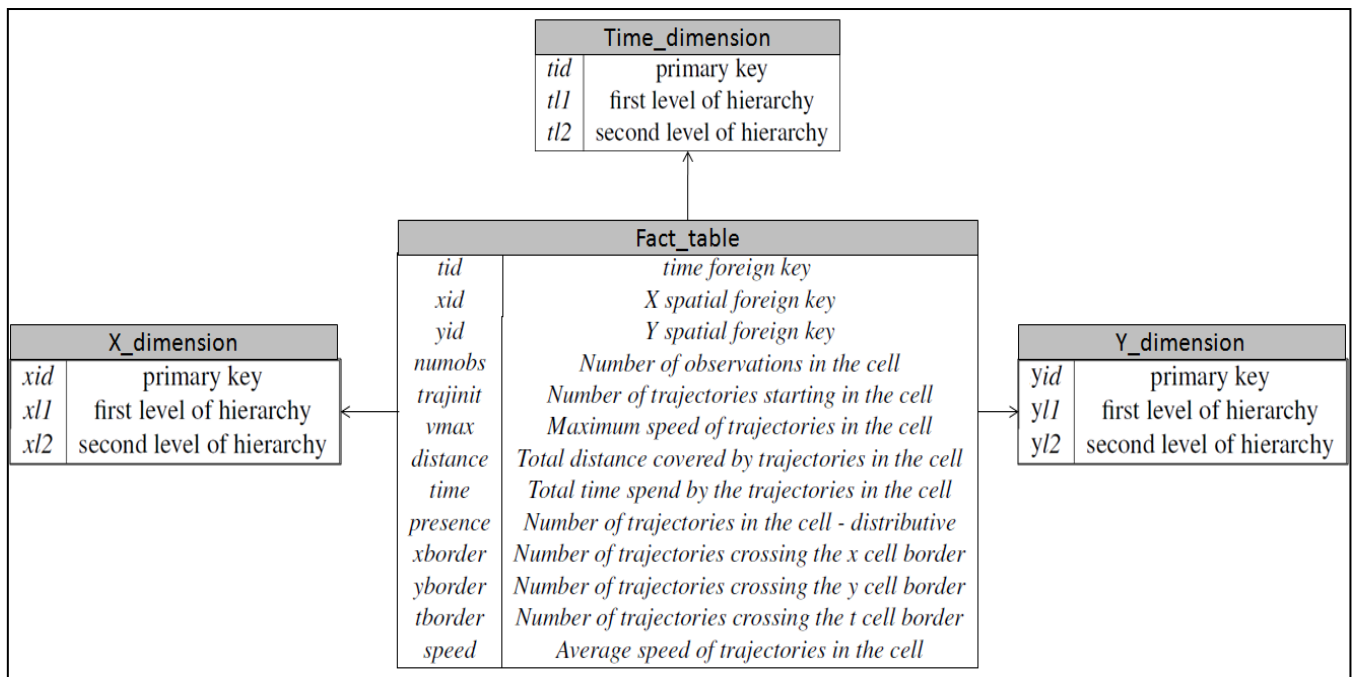


Figura A.2: Modelagem baseada em coordenadas espaciais (adaptação de [BRAZ07])

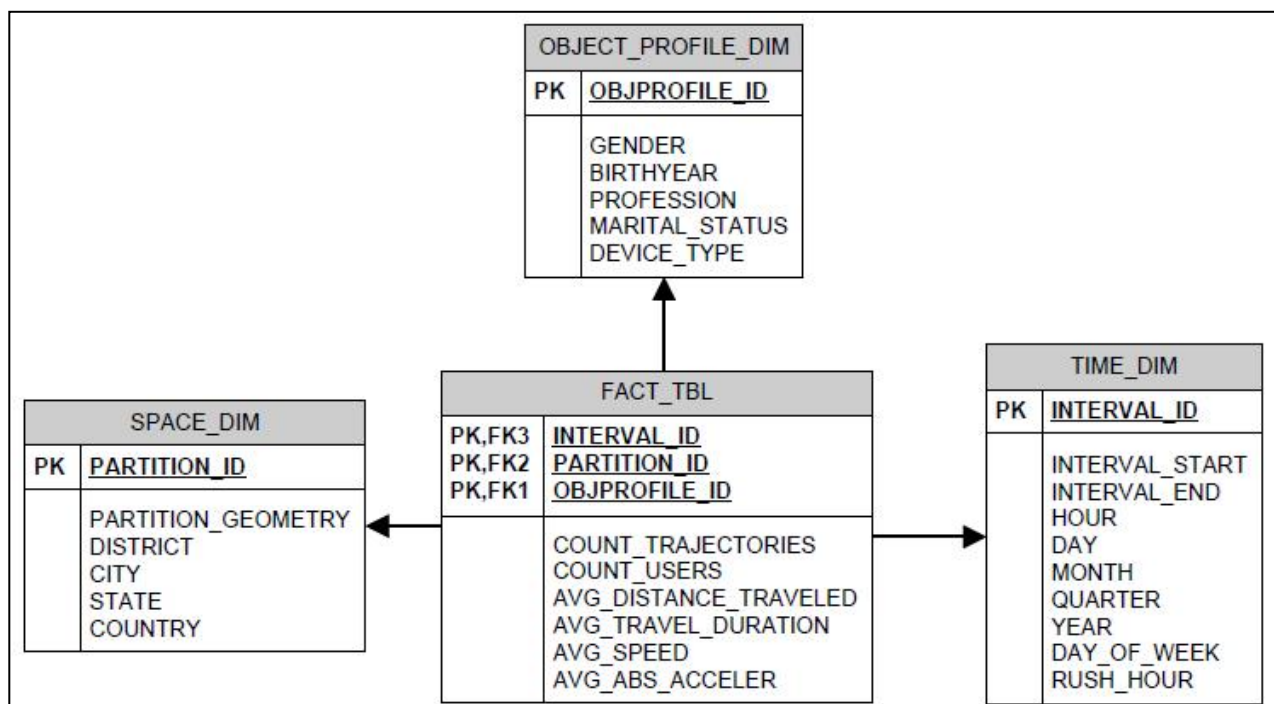


Figura A.3: Modelagem baseada nas informações dos objetos móveis ([MARK08])

Anexo B – Imagem do processo de ETL proposto

Novamente, por respeito à formatação do documento, a imagem que sistematiza o processo de ETL proposto para a ferramenta TETL na seção 2.4.1 foi exibida em tamanho reduzido. Para facilitar a visualização do processo, a Figura B.1 é colocada neste anexo.

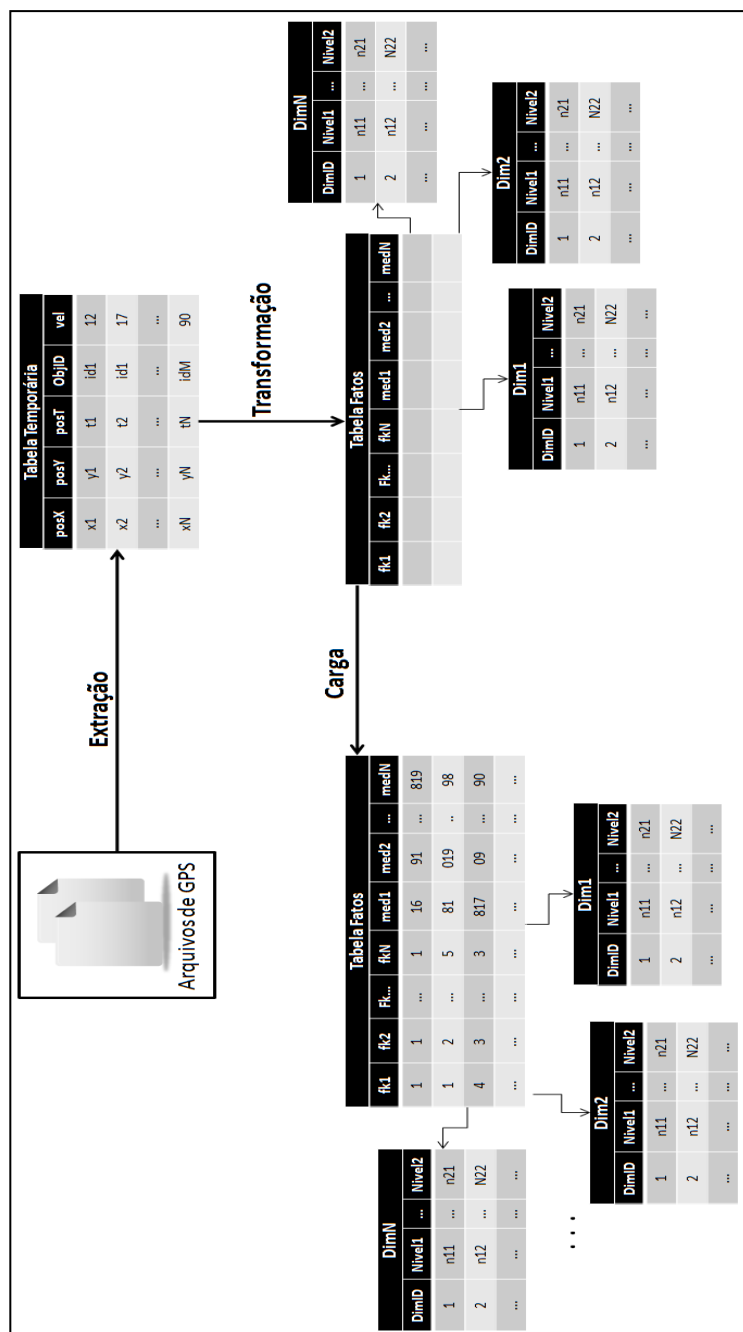


Figura B.1: Etapas do processo de ETL na ferramenta TETL

Assinaturas

Valéria Cesário Times (Orientadora)

Maria Carolina Torres da Silva (Aluna)