3
Sistemática Proposta

A sistemática proposta neste trabalho foi baseada nos conceitos e métodos revisados na literatura. A sistemática parte do desenho conceitual das estatísticas que é aplicado no processo de Dwing, transformando os microdados em formato adequado para análise e aberto. Posteriormente, essas informações podem ser disponibilizadas na Web para o uso de maneira rápida e barata.

A aplicação dessa sistemática para tornar as estatísticas públicas acessíveis via Web, justifica-se pelos os seguintes itens:

a) A Web é a plataforma utilizada pelo e-Governo e os órgãos oficiais de estatística migraram para a Web suas estruturas de disseminação de informações estatísticas, como se pode notar pelo texto retirado do site do IBGE[11]: "O IBGE estabelece, pela Internet seu principal canal de comunicação com o usuário disponibilizando os resultados das pesquisas".

b) Os dados disseminados em uma estrutura multidimensional na Web (WOLAP) oferecem mais recurso para consultas dos dados e dos conceitos da pesquisa, uma vez que as estatísticas públicas são disseminadas em formatos e recursos sem capacidade de análise. Desta forma, conseguir-se-á uma maior agilidade no acesso e uso das informações estatísticas, sem custo adicional por esse serviço.

c) O uso de ferramentas OLAP permite aos usuários obter uma visão multidimensional dos dados estatísticos e metadados. Essas visões multidimensionais auxiliam efetivamente aos analistas, pois eles podem verificar tendências nos dados, utilizando dados resumidos, trocando as dimensões de lugar e navegando através de suas hierarquias. Além disso, os analistas podem testar suas hipóteses e

[11] IBGE – Instituto Brasileiro de Geografia e Estatísticas. Instituição pública responsável pela produção, compilação e disseminação de estatísticas públicas e pela coordenação do Sistema Estatístico Nacional

pensar sobre questões que não haviam ainda sido levadas em consideração.

d) O armazenamento dos dados integrados em um DW torna as ferramentas OLAP mais eficazes, uma vez que os dados em um DW são integrados, limpos e transformados, garantindo assim uma maior qualidade e confiabilidade nos mesmos.

3.1. Cenário de Aplicação

O Sistema Estatístico Nacional (SEN) coordenado pelo IBGE[11] é composto por: censos demográficos; pesquisas amostrais e registros administrativos (Jannuzzi, 2006). Algumas das pesquisas que compõe o sistema são: Censos Demográficos, Censos populacionais, Pesquisa Nacional por Amostra de Domicílios (PNAD) e Pesquisa mensal de emprego(PME) , pesquisa de Emprego e Desemprego(PED) e Pesquisa de Orçamento Familiar (POF). Cada pesquisa é produzida de forma independente, com o uso de metodologias nas diversas fases de seu planejamento e execução. As metodologias são divulgadas explicitando os procedimentos usados e o amplo debate técnico proveniente de sua elaboração.

Na fase de planejamento, também é definido o instrumento de coleta dos dados, tratando-se de um questionário com perguntas fechadas. As perguntas do questionário fechado não são iguais às perguntas dos objetivos da pesquisa, mas estão intrinsecamente relacionadas com elas, de forma que a análise crítica das respostas às questões fechadas permite o estabelecimento de respostas aos objetivos da pesquisa.

O desenho conceitual de uma pesquisa estatística estabelece o âmbito da pesquisa, definindo a população investigada, os temas abordados, as variáveis, classificações adotadas e seus conceitos. Este trabalho de padronização visa oferecer uma comparabilidade entre os dados produzidos entre países e entre as pesquisas internas, além de delinear os conceitos abordados na pesquisa para o melhor entendimento dos seus resultados.

Assim, esta sistemática baseia-se na construção de um DW que será implementado na forma de um DM por vez, sendo os outros DMs implementados

de forma incremental (kimball, 2008). Dessa forma, a sistemática proposta deve ser aplicada em uma pesquisa por vez, permitindo que as pesquisas sejam integradas no DW em suas dimensões conformes, e a partir do DW e dos metadados, possa ser construído um cubo para cada pesquisa.

3.2. Formalização do Problema

Formalmente considera-se o seguinte problema:

Um conjunto de Estatísticas Públicas EP que são expressas em Questionários Fechados QF s.

Uma EP é dada por: EP={QF_k}, para k=1,2,3....,NQ, onde NQ é o numero de questionários aplicados na EP.

Para cada QF_k temos um conjunto de variáveis de análise V com as suas possíveis respostas ou categorias C.

Ou seja,

QF_k = {V_i, {C_{ij},i=1,2,...NV_k, j=1, 2,..., NCV_i} }, onde NV_k é o número de variáveis do formulário e NCV_i é o numero das possíveis respostas de V_i

Procura-se um modelo dimensional do banco de dados (um cubo) que será disponibilizado para análise.

Solução:

O modelo dimensional é dado por um esquema estrela com uma tabela fato Fato_QF_k central rodeada por várias tabelas de dimensão Dim_codV_i correspondentes as variáveis V_i , além da dimensão T da variável tempo.

Define-se a tabela fato por Fato_QF_k (LF, T, V_i, i=1,2,....NV_k, VM) onde LF representa as linhas da tabela fato e VM é uma variável de medida associada a pesquisa em questão.

Essa tabela é populada por {[lf, T_{lf}, Ci_{lf}, i=1,2,...NV_k , VM_{lf}]} onde C_{ilf}= valor respondido (categoria) da variável V_i numa linha do fato lf e VM_{lf} é o valor medido associado a essa linha do fato.

Para cada dimensão i (i=1,2,...NV_k):

Define-se o esquema da tabela dimensão por Dim_codV_i (C_i, nomeC_i).

Essa tabela é populada por {[C_{ij},descC_{ij}) para j=1,2,...NCV_i.

A sistemática proposta é ilustrada na Figura 7 e suas etapas serão detalhadas nos próximos tópicos. Para cada uma das etapas, é apresentado um produto final, que podem ser considerados resultados parciais da sistemática. A sistemática foi baseada no processo de Data Warehousing proposto por Kimball (1996)

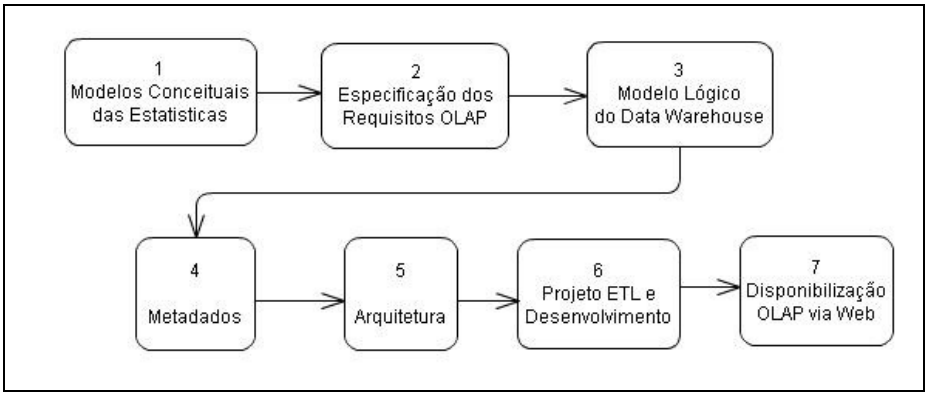

Figura 7 – Resumo da sistemática (fonte: a autora)

3.2.1. Etapa 1 – Modelos conceituais das estatísticas

Objetivo: Identificar nos microdados resultantes da estatística, quais as variáveis (dimensões) e medidas do contexto, definidos pela instituição pública na fase de elaboração de uma pesquisa, pontuando também as variáveis compatíveis entre as pesquisas já existentes no DW. Devem ser respondidas as seguintes questões: Qual a população investigada? Quais foram os elementos investigados? Quais variáveis foram observadas? O que foi medido?

Importância: No ponto de vista conceitual dimensional, essa sistemática trabalha com várias estrelas onde cada pesquisa pode ser integrada em suas dimensões conformes, por isso é nessa etapa que identifica-se os possíveis pontos de integração entre as pesquisas.

Como o objeto observado para ser modelado é uma estatística, a revisão de alguns conceitos básicos de estatística pode ajudar no entendimento da construção dos modelos criados nessa sistemática.

População: é o conjunto de elementos (p. ex., indivíduos, domicílios) que devem abranger o estudo e que são passíveis de serem observados, com respeito às características (variáveis) que se pretende levantar.

Amostra: É qualquer subconjunto da população que é utilizada no estudo, ou seja, é uma parte selecionada da totalidade das observações abrangidas pela população, através da qual se pode concluir sobre algumas características populacionais.

Variáveis: São as características que podem ser observadas (ou medidas) em cada elemento da população, sob as mesmas condições. A variável deve estar definida de tal forma que cada elemento observado tenha um – e apenas um – resultado (valor ou atributo) associado a essa variável. (Barbetta, 2006).

Desta forma, um questionário é aplicado aos elementos investigados na população que se transformam nos fatos e, as variáveis levantadas na pesquisa serão observadas sob vários pontos de vista, e esses pontos de vista são as dimensões de análise do modelo conceitual dimensional. Vale a pena observar que, nesta fase, todas as dimensões têm a mesma estrutura, conforme ilustrado na Figura 8. A granularidade da pesquisa deve ser identificada nessa etapa.

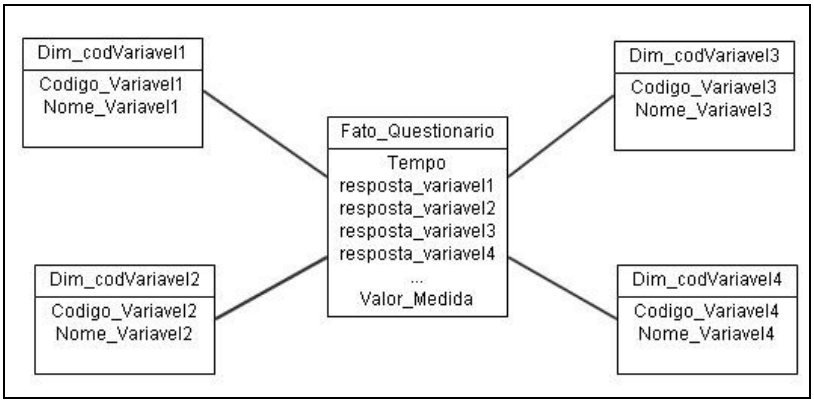

Figura 8 – Esquema dimensional conceitual de uma estatística pública

Passos:

- Identificar os elementos investigados pela estatística (fatos)
- Identificar a granularidade da pesquisa (grão)
- Identificar as dimensões a partir das variáveis de análise (cada pergunta realizada no questionário é uma variável de análise)
- Identificar as medidas a partir das variáveis de medidas
- Mapear as informações identificadas no modelo dimensional

Fechamento: o encerramento desta etapa é o esquema dimensional conceitual inicial do DM.

3.2.2. Etapa 2 - Especificação dos requisitos OLAP

Objetivo: Em geral, as variáveis não são analisadas conforme elas são coletadas. Dessa forma, o objetivo desta etapa é analisar a necessidade de acrescentar novos atributos de análise gerando assim novas hierarquizações para as dimensões. Por exemplo, no caso de uma variável idade, seria interessante acrescentar mais um atributo no esquema da dimensão indicando uma faixa de idade. Nesta etapa, devem ser respondidas as seguintes questões: Quais as possibilidades de análise de uma variável? Quais os níveis de detalhes da informação e quais são necessários?

Importância: A especificação de requisitos de análise OLAP é um dos pontos mais importantes da sistemática.Segundo Kimball (2008) a qualidade de um DW pode ser medida pela qualidade de suas dimensões, que são a porta de entrada do DW e a principal interface do usuário com os dados.

Geralmente, as pesquisas contêm muitas variáveis de análise, então, para facilitar essa etapa, as variáveis devem ser agrupadas de acordo com as necessidades de análise.

Por exemplo, as variáveis “renda per capita” ou “rendimento mensal” ou “rendimento de aposentadoria” armazenam valores em moeda e para que eles sejam comparáveis entre os anos é interessante acrescentar mais um nível de análise derivada do valor coletado. Assim, os valores derivados poderiam ser ajustados de acordo com os intervalos: “Até ¼ salário mínimo”, “Mais de ¼ até ½ salário”, “Mais de ½ até 1 salário”, “Mais de 1 até 2 salários”, “Mais de 2 até 3 salários”, “Mais de 3 até 5 salários” e “Mais de 5 salários”.

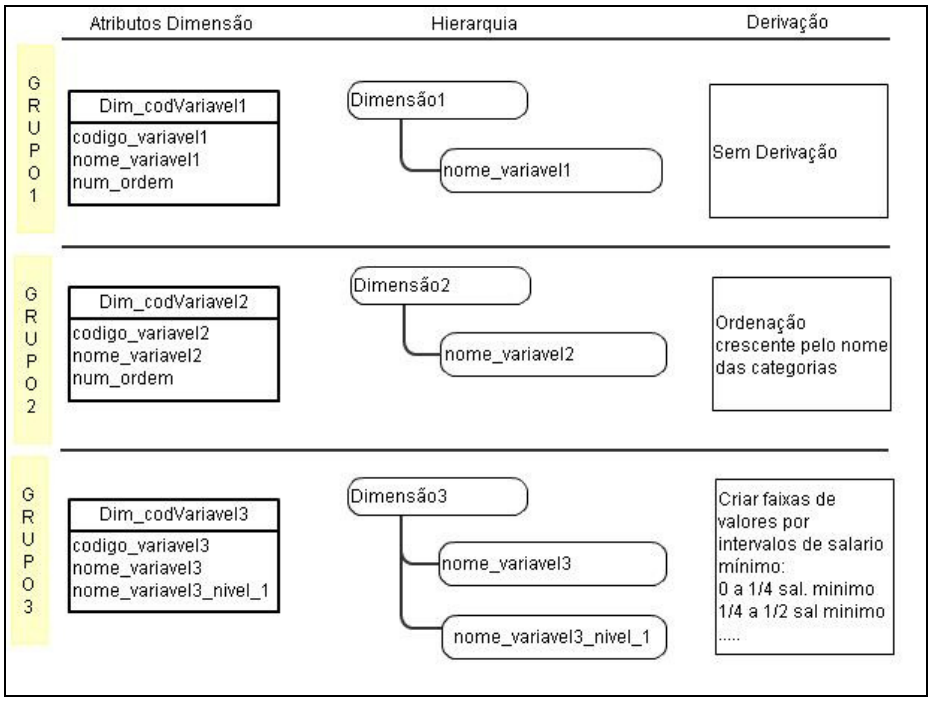

Figura 9 – Agrupamento de variáveis

Com as variáveis estatísticas classificadas em grupos, como no exemplo acima, deve-se ajustar a estrutura da dimensão correspondentemente. E para cada grupo devem ser descritas as regras de derivação. Esta etapa apresenta os passos descritos a seguir.

Passos:

- Agrupar as dimensões de acordo com a necessidade de hierarquia de análise
- Criar os modelos de dimensões padrões para atender os requisitos de análise
- Especificar as regras de derivação dos grupos
- Classificar as variáveis de medida quanto a aditividade, conforme descrito na seção 2.5.

Fechamento: esta etapa se encerra com a modificação do esquema das dimensões incluindo novos atributos necessários para o enriquecimento das análises. Para cada novo atributo acrescentado deve ser definida a regra de derivação de valores que represente a variável coletada (faixa, intervalos). Esse cálculo derivado pode ser automaticamente gerado analisando os dados da fonte na hora da extração.

3.2.3. Etapa 3 – Esquema lógico do DW/DM

Objetivo: Construir um esquema lógico do DW, com base no esquema conceitual da etapa 1 e nos esquemas hierárquicos das dimensões definidos na etapa 2, considerando os limites impostos pela tecnologia de banco de dados relacionais.

Importância: Concluir a construção do esquema dimensional lógico tornando-o disponível para a implementação do esquema físico.

Passos:

- Adequar a nomenclatura
- Criar as chaves primárias e estrangeiras
- Documentar as dimensões, fatod, atributos e medidas

Um esquema Lógico dimensional do DW com duas pesquisas é mostrado na Figura 10.

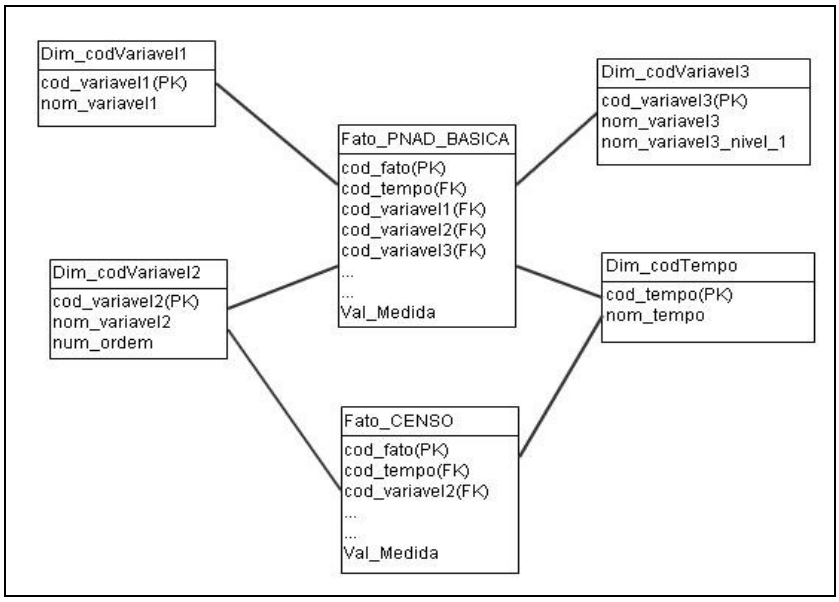

Figura 10 – Exemplo de esquema lógico dimensional

Fechamento: A etapa se encerra com a conclusão do esquema lógico do DW, utilizando o modelo relacional estrelado.

3.2.4. Etapa 4 - Metadados

Objetivo: Definir as tabelas complementares para armazenar os metadados técnicos e semânticos. Os metadados técnicos vão descrever as variáveis coletadas na pesquisa (estrutura) com suas categorias correspondentes. Os metadados semânticos armazenam descrições orientadas ao conteúdo da pesquisa com maior nível de detalhamento (definidos pelo instituto de pesquisa), estes conteúdos serão levados para os cubos.

Importância: Todos os componentes de um processo de DWing devem ser administrados a partir de um repositório de metadados (Quix, 1999). Além disso, conforme citado no capítulo 2, as ferramentas OLAP que suportam o XMLA oferecem métodos de consultas aos seus metadados. Dessa forma, podem ser apresentadas aos usuários finais várias visões de dados orientadas a uma pesquisa estatística e também as informações no nível de conceitos da pesquisa.

Os microdados disponibilizados na Web pelas instituições de pesquisa, conforme a seção 2.4, contém quase todas as informações necessárias para a aplicação dessa sistemática. A única informação que deve ser acrescentada pela sistematica é a classificação quanto ao tipo de variável (definida na etapa 2).

Passos:

- Construir os esquemas conceitual e lógico para armazenamento dos metadados

Fechamento: A etapa se encerra com a criação de uma instância dos metadados.

3.2.5. Etapa 5 – A Arquitetura

Objetivo: o objetivo principal dessa etapa é definir o conjunto de tecnologias necessárias para cada fase do projeto de DW. Nessa etapa deve-se responder a seguinte questão: Como será feito?

Importância: A escolha das ferramentas, dos utilitários e das plataformas deve ser cuidadosa e estar alinhada ao objetivo do projeto, pois se pode fazer uso de soluções caras e complexas.

O requisito crítico da arquitetura proposta é o servidor OLAP. Deve-se escolher um servidor que execute queries MDX e que funcione como XMLA provider, obedecendo assim às especificações do XMLA Council. A arquitetura proposta é representada na Figura 11.

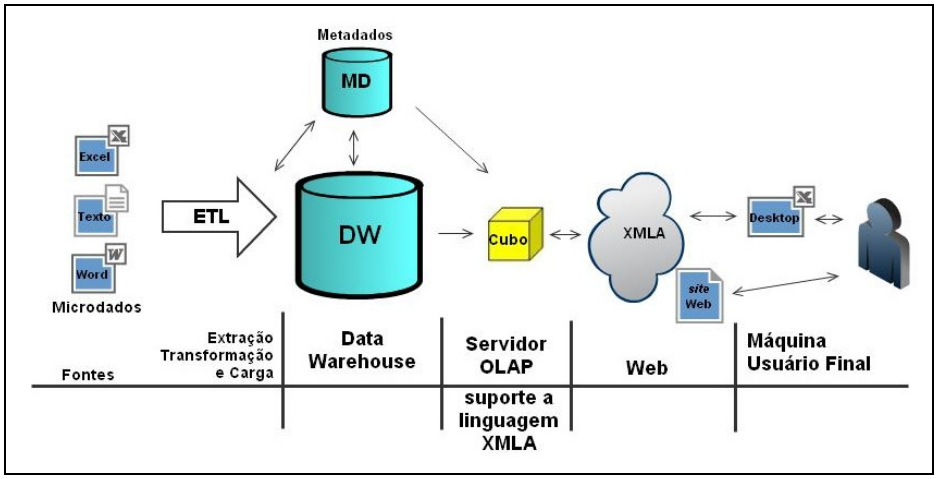

Figura 11 - Arquitetura da sistemática proposta

Passos:

- Identificar o formato de disponibilização das fontes de dados
- Definir as ferramentas para o desenvolvimento do ETL
- Definir o repositório para os Metadados e Data Warehouse
- Definir o servidor OLAP
- Configurar e testar todos os elementos da arquitetura

Fechamento: o encerramento desta etapa apresenta um diagrama da arquitetura com os detalhes dos seus componentes.

3.2.6. Etapa 6 – Projeto de ETL e Desenvolvimento

Objetivo: Projetar as regras de construção do esquema físico do DW e do ETL para que o fluxo de dados ocorra. O fluxo inicia no acesso às fontes, seguido pelo processo de limpeza, transformação e carga no DW, até chegar no destino final que é cubo no servidor OLAP. Todo o procedimento deve ser realizado partindo dos metadados (semânticos e técnicos).

Importância: Essa etapa é crítica, pois são os processos de ETL definidos aqui que garantem a qualidade dos dados apresentados para os usuários. Essa etapa deve ser desenvolvida centrada nos metadados da pesquisa estatística.

No processo de ETL, muitas regras devem ser estabelecidas e para facilitar a ordem de execução desse processo ele foi dividido em categorias. A categoria de construção do esquema físico do DW e construção do cubo devem

ser executadas apenas como carga inicial, os demais passos devem ser executados para realização das cargas incrementais do DW.

Passos:

Preparação das fontes

- Acessar e identificar as fontes de dados
- Padronizar os arquivos de metadados da pesquisa
- Construir o esquema físico dos metadados
- Importar os microdados da pesquisa para os metadados do DW

Construção do esquema físico do DW

- Criar a estrutura básica das tabelas dimensões
- Criar a tabela fato

Limpeza e transformação na área temporária

- Realizar a carga da tabela fato
- Realizar a carga das tabelas dimensões
- Definir cenários para validar a integridade dos dados
- Incluir as PK nas tabelas dimensões
- Incluir as FK na tabela fato
- Alterar as tabelas dimensões incluindo as hierarquias de análise necessárias para cada tipo de dimensão
- Executar a derivação de valores necessária para cada tipo de dimensão

Construção do Cubo

- Criar as dimensões do cubo
- Criar o cubo
- Processar o cubo

Cenários para validar as informações analíticas geradas

- Definir os cenários para validar a informação gerada comparando com as informações publicadas pelo instituto responsável pela produção da estatística

Fechamento: a execução bem sucedida da seqüência de passos descrita nessa etapa visa garantir a qualidade dos dados, tendo como o produto final um cubo de dados com as informações organizadas de forma analíticas.

3.2.7. Etapa 7 – Disponibilizar os dados na Web

Objetivo: Disponibilizar os dados na Web para serem utilizados via uma aplicação WOLAP.

Importância: Garantir o acesso democrático às estatísticas públicas, entregando para o usuário final um modelo que permita o entendimento e a exploração de um grande volume de dados com alta flexibilidade e performance.

Passos:

- Configurar o servidor de aplicações Web para habilitar o serviço de acesso OLAP

Fechamento: O principal produto dessa etapa é o cubo acessível via web, sem custo.

No próximo capítulo, será descrita a aplicação da sistemática proposta. Para sua validação, como prova de conceito serão utilizados dados governamentais reais.