AQUAWARE: Um Ambiente de Suporte à Qualidade de Dados em *Data Warehouse*

Glenda Carla Moura Amaral

Universidade Federal do Rio de Janeiro
Instituto de Matemática
Núcleo de Computação Eletrônica
Mestrado

Maria Luiza Machado Campos, PhD

Rio de Janeiro 2003

AQUAWARE: Um Ambiente de Suporte à Qualidade de Dados em Data *Warehouse*

Glenda Carla Moura Amaral

Dissertação submetida ao corpo docente do Instituto de Matemática – Universidade Federal do Rio de Janeiro – UFRJ, como parte dos requisitos necessários à obtenção do grau de Mestre.

Aprovada por:	
Prof ^a . Maria Luiza Machado Campos – PhD.	_Orientador
Paulo oh ppnds).	_
Prof. Paulo Figueiredo Pires – D.Sc.	
Prof. Jano Moreira de Souza PhD	
Prof. Marcus Costa Sampaio – PhD.	-

Rio de Janeiro 2003

FICHA CATALOGRÁFICA

Amaral, Glenda Carla Moura.

AQUAWARE: Um Ambiente de Suporte à Qualidade de Dados em *Data Warehouse |* Glenda Carla Moura Amaral – Rio de Janeiro, 2003.

xv, 164f.: il.

Dissertação (Mestrado em Informática) — Universidade Federal do Rio de Janeiro — UFRJ, Instituto de Matemática — IM, Núcleo de Computação Eletrônica — NCE, 2003.

Orientadora: Maria Luiza Machado Campos

1. Qualidade de Dados. 2. Data Warehouse Federado. 3. Data Warehouse Interorganizacional. 4. Métricas de Qualidade. 5. Metadados de Qualidade.

A Deus, pela oportunidade de desenvolver este trabalho e ao meu irmão Júnior.

AGRADECIMENTOS

A Deus, por tudo...

Aos meus pais, Carlos Alberto e Naly pelo apoio em todas as minhas escolhas e iniciativas e por nunca terem medido esforços para me dar a melhor formação possível.

Ao meu pequeno irmão Júnior, pelo carinho e compreensão, evitando fazer bagunça em nosso quarto durante os meus momentos de estudo.

À Professora Maria Luiza, uma pessoa admirável e inteligente, pela orientação e amizade em todos os momentos, pelo conhecimento e incentivo transmitidos, por estar sempre disponível para todas as dúvidas.

À Marinha do Brasil, através da Diretoria de Administração da Marinha, pela oportunidade que me foi concedida, para a realização deste curso.

Aos Almirantes Barros e Ostwald e aos Comandantes Alevatto, Moreira, Rodrigues Neto, Magno, Barros e Arnaldo, pelo apoio no sentido de propiciar o meu licenciamento parcial para a realização deste curso.

Ao Comandante Hamilton, por ter me incentivado a ingressar neste curso e pelo exemplo constante na busca do conhecimento.

Ao PRODERJ, em especial ao Maurício Abreu, por nos concederem o privilégio de utilizarmos o Projeto do Portal de Informações Consolidadas de Governo, como exemplo de aplicação nesta pesquisa.

Aos consultores Marcelo e Luís da Hiper e ao Fernando Campos da Coca-Cola, pela atenção e conhecimentos transmitidos sobre suas experiências em projetos envolvendo a tecnologia de *Data Warehouse*.

A Kelli, Sandro e Luciano, pela dedicação no projeto de implementação do protótipo do AQUAWARE.

Aos Professores Marcus Sampaio, Jano Moreira e Paulo Pires, pela gentileza de aceitar o convite para participar da avaliação deste trabalho.

Finalmente, a todos aqueles que direta ou indiretamente contribuíram na elaboração deste trabalho.

RESUMO

AMARAL, Glenda Carla Moura. **AQUAWARE: Um Ambiente de Suporte à Qualidade de Dados em** *Data Warehouse*. Orientadora: Maria Luiza Machado Campos. Rio de Janeiro: UFRJ/IM, 2003. Dissertação (Mestrado em Informática).

No contexto econômico atual, onde a competição entre as empresas está cada vez mais relacionada à sua capacidade de transformar informação em conhecimento e este último em decisões e ações de negócio, o Data Warehouse (DW) exerce um papel fundamental como ferramenta de apoio aos processos de tomada de decisão. Contudo, para que se possa tirar vantagem dos recursos de informação de forma satisfatória, é necessário que as informações utilizadas sejam confiáveis, ou que pelo menos, o grau de confiabilidade das mesmas possa ser adequadamente calculado e considerado, durante o processo de apoio à decisão. Nesse processo, a qualidade de dados funciona como um instrumento de avaliação e garantia da confiabilidade. Para viabilizar o fornecimento de informações sobre a qualidade dos dados do DW para o usuário final, é caracterizado nesta dissertação um ambiente computacional que tem como objetivo atuar como infraestrutura para o fornecimento de informações de qualidade para ferramentas que atuem como clientes do DW, denominado Ambiente de Suporte à Qualidade de dados em Data Warehouse (AQUAWARE). Como parte integrante desta solução foram definidos indicadores para a avaliação da qualidade dos dados, sob a perspectiva do usuário final. O AQUAWARE representa um avanço em relação às consultas analíticas tradicionais, pois provê subsídios que permitem ao usuário avaliar até que ponto é seguro apoiar-se nas informações apresentadas, conferindo mais confiabilidade ao processo de tomada de decisão.

ABSTRACT

AMARAL, Glenda Carla Moura. **AQUAWARE: Um Ambiente de Suporte à Qualidade de Dados em** *Data Warehouse*. Orientadora: Maria Luiza Machado Campos. Rio de Janeiro: UFRJ/IM, 2003. Dissertação (Mestrado em Informática).

In the current economy the Data Warehouse constitutes a fundamental tool to support decision making processes. Companies now compete on the ability to turn information into knowledge and this one into decisions and business actions. However, to better take advantage of enterprise information assets, it is necessary that information to be reliable, or at least, its reliability degree can be appropriately obtained and considered during decision making. In this process, data quality management assumes special importance, as it is both complex and essential. In order to make DW quality data available to the end user, this dissertation characterizes a computational environment named Data Warehouse Data Quality Support Environment (AQUAWARE) to be used as infrastructure for supplying quality information to DW client tools. In addition, we define some data quality indicators, used to evaluate quality under the end user's perspective. Queries performed through AQUAWARE represents an advance when compared to traditional analytical queries, as it enables the end user to evaluate to what extent it is safe to trust in the presented information, providing more reliability to the decision making process.

LISTA DE SIGLAS

API - Application Program Interface

AQUAWARE - Ambiente de Suporte à Qualidade de Dados em Data Warehouse

BD – Banco de Dados

CASE - Computer-Aided Software Engineering

CIF - Corporate Information Factory

CWM - Common Warehouse Metamodel

DTD – Document Type Definition

DW – Data Warehouse

DWQ - Data Warehouse Quality

EDW - Enterprise Data Warehouse

ERP - Enterprise Resource Planning

ETL - Extraction, Transformation and Loading

GQM - Goal Question Metric

HTTP – Hypertext Transfer Protocol

IRDS - Information Resouce Dictionary System

ISO-International Standardization Organization

J2EE – Java 2 Enterprise Edition

JMI – Java Metadata Interface

JOLAP - Java OLAP Interface

MDX - Multidimensional Expressions

MOF - Meta Objects Facility

NASA – National Aeronautics and Space Administration

ODBC - Open DataBase Connection

ODS - Operational Data Store

OLAP - On-line Analytical Processing

OLTP – On-line Transactional Processing

OMG - Object Management Group

PRODERJ - Centro de Processamento de Dados do Estado do Rio de Janeiro

RMQ - Repositório de Metadados da Qualidade

SAW - Simple Additive Weighting

SECTI - Secretaria de Ciência, Tecnologia e Inovação

SQL – Structured Query Language

TDWI - The Data Warehouse Institute

UML - Unified Modeling Language

VV&C - Verificação, Validação e Certificação

XMI - XML Metadata Interface

XML - eXtensible Markup Language

XMLA - XML for Analysis

LISTA DE ILUSTRAÇÕES

Figura 2.1 – Objetivos, Questionamentos e Métricas (BASILI et al., 1994)	24
Figura 2.2 - Qualidade de Dados na CIF (INMON, 2003)	40
Figura 2.3 - Modelagem Dimensional da Auditoria da Qualidade de Dados	42
Figura 2.4 - Certificação do Data Warehouse (LOSHIN, 2001)	44
Figura 2.5 - Ligação entre fatores de qualidade e atividades do DW (JARKE e	
VASSILIOU, 1997)	47
Figura 2.6 - Metamodelo de qualidade DWQ simplificado (JEUSFELD et al.,1998	3) 49
Figura 2.7 – O Metamodelo da Qualidade de Vassiliadis	51
Figura 2.8 – A classe MEASUREMENT do CWM (OMG, 2003)	55
Figura 3.1 – A Arquitetura de DW Federado proposta por Anand (1997)	60
Figura 3.2 – A Arquitetura de DW Federado proposta por Mangisengi e Ruber (20	001)61
Figura 3.3 – Interoperabilidade em DWs Federados usando XML	
(MANGISENGI,HUBER,2001)	63
Figura 3.4 – Exemplo de arquitetura usando XMLA	65
Figura 3.5 – O modelo JOLAP (HYPERION, SUN, 2003)	66
Figura 4.1 – Arquitetura de Quatro Níveis do IRDS	70
Figura 4.2 – Esquema simplificado da solução proposta por Campos (1993)	72
Figura 4.3 – A Arquitetura do AQUAWARE	73
Figura 4.4 – Integração entre o Metamodelo do RMQ e o CWM	75
Figura 4.5 - Interação entre o cliente, o Serviço de Consulta e o Serviço de Qualid	dade76
Figura 4.6 – O AQUAWARE em DW federado de contexto interorganizacional	78
Figura 4.7 – Necessidades de representação no Metamodelo do RMQ	80
Figura 4.8 – O Metamodelo do Repositório de Metadados da Qualidade	81
Figura 4.9 – Metamodelo do RMQ sob a perspectiva do usuário final	85
Figura 4.10 – Modelo Dimensional do Cubo das Dimensões	87
Figura 4.11 – Modelo Dimensional do Cubo dos Fatos	88
Figura 4.12 - Modelo Dimensional de um Cubo de Tuplas	89
Figura 4.13 – Modelo Dimensional de um Cubo de Células	90
Figura 4.14 – Modelo Dimensional do Cubo de Anotações das Dimensões	91
Figura 4.15 – Modelo Dimensional do Cubo de Anotações dos Fatos	91

Figura 4.16 – Modelo Dimensional do Cubo de Anotações de Tuplas	91
Figura 4.17 – Modelo Dimensional do Cubo de Anotações de Células	92
Figura 4.18 – Arquitetura do AQUAWARE utilizando XMLA	97
Figura 6.1 – Arquitetura do Ambiente de <i>Data Warehoue</i> do PRODERJ	111
Figura 6.2 – Cubo das Dimensões (PRODERJ)	112
Figura 6.3 – Cubo dos Fatos (PRODERJ)	113
Figura 6.4 – Cubo de Anotações das Dimensões (PRODERJ)	113
Figura 6.5 – Cubo de Anotações dos Fatos (PRODERJ)	113
Figura 6.6 – Diagrama de Atividades da Dinâmica de Atuação dos Componentes	114

LISTA DE TABELAS

Tabela 2.1 - Características e Subcaracterísticas da Qualidade de <i>Software</i> (ISO 91	26,
1991)	22
Tabela 2.2 - Critérios da qualidade de dados nas visões interna e externa (WAND e	
WANG, 1996)	28
Tabela 2.3 - Agrupamento de Critérios da Qualidade de Dados (WANG et al., 1990)	
Tabela 2.4 – Critérios e categorias da qualidade de dados, segundo Loshin (2001)	29
Tabela 2.5 – Quadro de Referência das propostas de critérios para a qualidade de d	lados
	32
Tabela 2.6 - Comparação entra as abordagens para a qualidade de dados no DW	53
Tabela 3.1 - Comparação XMLA x JOLAP	68
Tabela 4.1 – Padronização da terminologia para os níveis da arquitetura multi-cam	adas
	71
Tabela 4.2 – Componentes para a carga dos Cubos de Qualidade	
Tabela 5.1 – Indicadores para os Fatos	106
Tabela 5.2 – Indicadores para as Dimensões	108
Tabela 6.1 – Qualidade das Dimensões	116
Tabela 6.2 – Qualidade dos Fatos em relação à dimensão MUNICIPIO	116

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO	16
1.1 MOTIVAÇÃO	16
1.2 HIPÓTESE	18
1.3 OBJETIVO	18
1.4 ESTRUTURA	19
CAPÍTULO 2 - QUALIDADE DE DADOS	21
2.1 QUALIDADE DE SOFTWARE	21
2.1.1 O PARADIGMA GOAL QUESTION METRIC	24
2.2 TIPOS DE AVALIAÇÃO DA QUALIDADE DE DADOS EM BANCOS DE	
DADOS	25
2.3 ELEMENTOS PARA A DESCRIÇÃO DA QUALIDADE DE DADOS	26
2.4 PROPOSTAS DE CRITÉRIOS PARA A AVALIAÇÃO DA QUALIDADE D	ЭE
DADOS	27
2.5 ABORDAGENS PARA A QUALIDADE DE DADOS EM DATA WAREHOU	USE 34
2.5.1 A ABORDAGEM DE DAVID MARCO	35
2.5.2 A ABORDAGEM DE JEFF ROTHENBERG	36
2.5.3 A ABORDAGEM DE INMON	39
2.5.4 A ABORDAGEM DE KIMBALL	41
2.5.5 A ABORDAGEM DE LOSHIN	43
2.5.6 A ABORDAGEM DE LARRY ENGLISH	45
2.5.7 A ABORDAGEM DO DWQ	47
2.5.8 A ABORDAGEM DE VASSILIADIS	50
2.5.9 COMPARAÇÃO DAS ABORDAGENS APRESENTADAS	52
2.6 QUALIDADE DE DADOS E O <i>COMMON WAREHOUSE METAMODEL</i>	54
CAPÍTULO 3 - <i>DATA WAREHOUSES</i> FEDERADOS E O CONTEXTO	
INTERORGANIZACIONAL	57
3.1 DATA WAREHOUSE EM CONTEXTO INTERORGANIZACIONAL	57
3.2 DATA WAREHOUSES FEDERADOS	59
3.2.1 A ARQUITETURA PROPOSTA POR ANAND	59
3.2.2 A ARQUITETURA PROPOSTA POR MANGISENGI E HUBER	60

3.2.3 CONSIDERAÇÕES SOBRE O ESQUEMA FEDERADO	62
3.3 INTEROPERABILIDADE EM AMBIENTE OLAP	62
3.3.1 INTEROPERABILIDADE NA ARQUITETURA DE MANGISENGI E HUB	ER
	63
3.3.2 XML FOR ANALYSIS	64
3.3.3 JAVA OLAP INTERFACE	65
3.3.4 COMPARAÇÃO DAS PROPOSTAS DE INTEROPERABILIDADE	67
CAPÍTULO 4 - AMBIENTE DE SUPORTE À QUALIDADE DE DADOS EM	
DATA WAREHOUSE – AQUAWARE	69
4.1 ARQUITETURA DE MODELAGEM MULTI-CAMADAS	69
4.2. IDENTIFICAÇÃO DE OBJETOS NA ARQUITETURA DE MODELAGEM	
MULTI-CAMADAS	71
4.3 CARACTERÍSTICAS DO AMBIENTE DE SUPORTE À QUALIDADE DE	
DADOS EM DATA WAREHOUSE (AQUAWARE)	72
4.4 O AQUAWARE EM <i>DW</i> FEDERADO INTERORGANIZACIONAL	76
4.5 O METAMODELO DO REPOSITÓRIO DE METADADOS DA QUALIDADE	E 79
4.5.1 NECESSIDADES DE REPRESENTAÇÃO NO METAMODELO	
4.5.2 DESCRIÇÃO DO METAMODELO	80
4.5.3 METAMODELO DO RMQ SOB A PERSPECTIVA DO USUÁRIO FINAL .	84
4.6 CUBOS DE QUALIDADE	86
4.6.1 O CUBO DAS DIMENSÕES	87
4.6.2 O CUBO DOS FATOS	87
4.6.3 OS CUBOS DE TUPLAS	88
4.6.4 OS CUBOS DE CÉLULAS	89
4.6.5 OS CUBOS DE ANOTAÇÕES	90
4.7 BIBLIOTECA DE COMPONENTES	93
4.7.1 CARGA DO MODELO DE QUALIDADE NO RMQ	93
4.7.2 GERAÇÃO DAS MEDIDAS DE QUALIDADE	93
4.7.3 CÁLCULO DOS OBJETIVOS	94
4.7.4 CARGA DOS CUBOS DE QUALIDADE	95
4.7.5 OS SERVIÇOS DE CONSULTA E QUALIDADE	96

CAPÍTULO 5 - INDICADORES DA QUALIDADE DE DADOS NO DW SOB A
PERSPETIVA DO USUÁRIO FINAL
5.1 SELEÇÃO DOS CRITÉRIOS DA QUALIDADE
5.2 DEFINIÇÃO DOS INDICADORES (MÉTRICAS)
5.2.1 INDICADORES PARA OS FATOS
5.2.2 INDICADORES PARA AS DIMENSÕES
5.3 – CONSIDERAÇÕES SOBRE A DEFINIÇÃO DE INDICADORES 108
CAPÍTULO 6 - EXEMPLO DE APLICAÇÃO: CENTRO DE
PROCESSAMENTO DE DADOS DO ESTADO DO RIO DE JANEIRO
(PRODERJ)
6.1 CENÁRIO: PRODERJ 109
6.2 ARQUITETURA DO AMBIENTE DE <i>DATA WAREHOUSE</i> DO PRODERJ 110
6.3 OS CUBOS DE QUALIDADE
6.4 DINÂMICA DE ATUAÇÃO DA BIBLIOTECA DE COMPONENTES 114
6.5 UMA CONSULTA À QUALIDADE DE DADOS NO <i>DATA MART</i>
ESTRATÉGICO DO GOVERNO ESTADO DO RIO DE JANEIRO115
6.6 CONSIDERAÇÕES SOBRE A CONSULTA DE QUALIDADE 117
CAPÍTULO 7 - CONCLUSÕES E TRABALHOS FUTUROS 118
BIBLIOGRAFIA
ANEXO I - DESCRIÇÃO DOS MÉTODOS DISCOVER E EXECUTE DO XMLA
ANEXO I I - DESCRIÇÃO DOS CRITÉRIOS DE QUALIDADE DA TABELA
2.5
ANEXO III - LISTA DOS COMPONENTES DA BIBLIOTECA 138
ANEXO IV - ESPECIFICAÇÃO DOS COMPONENTES DA BIBLIOTECA 139
ANEXO V - ESPECIFICAÇÃO DOS MÉTODOS DISCOVER E EXECUTE DO
XMLA-Q
ANEXO VI - COMANDOS XMLA DA CONSULTA DO EXEMPLO DE
APLICAÇÃO155

CAPÍTULO 1

INTRODUÇÃO

1.1 MOTIVAÇÃO

Nas economias de natureza agrícola, a terra e a mão-de-obra eram os fatores críticos para determinar o sucesso econômico. O capital e a tecnologia (arado, por exemplo) eram importantes, mas a comunidade agrícola podia subsistir com um mínimo de dinheiro e tecnologia, porém não sem terra e mão-de-obra. Com a revolução industrial, a tecnologia ganha importância, mas o capital e o trabalho passam a ser as forças motrizes do desenvolvimento econômico (CAVALCANTI, GOMES, 2000).

Uma nova realidade, no entanto, vem se impondo no cenário mundial: a chamada Sociedade do Conhecimento. Peter Drucker, já em 1968, enfatizava que "o conhecimento tornou-se o recurso essencial da economia". A economia do conhecimento desloca o eixo da riqueza e do desenvolvimento de setores industriais tradicionais (intensivos em mão-de-obra, matéria-prima e capital) para setores cujos produtos, processos e serviços são intensivos em tecnologia e conhecimento. Nela, a competição é cada vez mais baseada na capacidade de transformar informação em conhecimento e conhecimento em decisões e ações de negócio (CAVALCANTI, GOMES, 2000).

O surgimento de uma economia baseada no conhecimento modificou profundamente as exigências de informação das organizações no mundo inteiro. A obtenção de vantagem competitiva está diretamente relacionada à agilidade nas tomadas de decisão, que são feitas com base nos recursos de informação disponíveis nas organizações. A necessidade de análises mais sofisticadas e sínteses mais rápidas aumenta a cada dia. A prosperidade das empresas está diretamente relacionada à sua habilidade de obter, analisar e sintetizar informações usando seus sistemas de informação.

Para atender a estas novas necessidades, um novo conjunto de conceitos e ferramentas vem ganhando destaque nos últimos anos: a tecnologia de *Data Warehouse*¹ e suporte ao processamento analítico, que oferecem às organizações uma maneira flexível e eficiente de obter as informações necessárias a seus processos decisórios.

A crescente utilização do *Data Warehouse* (DW) em tomadas de decisão estratégicas, críticas para o negócio da empresa, aumenta a importância de apoiar-se em informações válidas, confiáveis e de qualidade. Da mesma forma que uma casa construída sobre alicerces fracos tende a desmoronar, conclusões baseadas em dados incorretos dificilmente irão produzir os resultados esperados. Dados de má qualidade podem ocasionar falhas nos processos de negócio e custos relacionados a pessoas, materiais, tempo, dinheiro e até mesmo a perda definitiva de clientes. O TDWI (*The Data Warehouse Institute*) estima que problemas relacionados à qualidade de dados custam, para os Estados Unidos, mais de \$600 bilhões por ano (ECKERSON, 2002).

O custo da má qualidade de dados pode ser exemplificado pelo caso de uma companhia de seguros que recebe cerca de 2 milhões de pedidos por mês, onde cada pedido contém 377 elementos de dados. Supondo que haja uma taxa de erros de 0,001, os dados dos pedidos conterão mais de 754.000 erros por mês e mais de 9,04 milhões de erros por ano. Se essa companhia considera que 10% dos elementos de dados são críticos para suas decisões e processos de negócio, será necessário corrigir quase 1 milhão de erros por ano. Se essa empresa estima um custo de \$10 por erro, o risco a que ela está exposta, em função da má qualidade de dados, é de \$10 milhões por ano (ECKERSON, 2002).

Apesar disso, decisões baseadas em informações resultantes de consultas analíticas são tomadas diariamente, nas empresas, sem que haja um conhecimento prévio do grau de qualidade dos dados envolvidos, aumentando o risco da produção de resultados inesperados, em função da má qualidade.

_

¹ No presente trabalho o termo *Data Warehouse* é utilizado para representar tanto o ambiente de DW quanto o banco de dados propriamente dito, dependendo do contexto.

No ambiente de *Data Warehouse* de uma organização, a obtenção de indicadores sobre a qualidade dos dados é uma tarefa complexa, que envolve dados provenientes de diversas fontes, capturados em momentos diferentes e por processos distintos.

Muitas vezes, parte das informações de origem é gerada fora da organização, havendo uma mesclagem de dados externos e internos, compondo um ambiente de *Data Warehouse* interorganizacional. Expor diferentes ambientes analíticos e possibilitar a troca de informações entre eles pode não ser trivial devido à sua heterogeneidade lógica, conceitual e física. Uma abordagem bastante utilizada no tratamento dos problemas de interoperabilidade entre *Data Warehouses* heterogêneos (plataformas distintas), autônomos e distribuídos é a tecnologia de ambientes federados.

Quanto aos tomadores de decisão, num contexto interorganizacional, é ainda mais difícil avaliarem o quanto podem confiar nos resultados das consultas analíticas, pois como parte dos dados são originados fora da organização, eles não têm como prever o grau de qualidade dos mesmos.

1.2 HIPÓTESE

No processo de tomada de decisão, a qualidade dos dados afeta o resultado das análises. Se o grau de qualidade das informações analíticas for adequadamente calculado e considerado durante o processo de apoio à tomada de decisão, é mais provável que as resoluções produzam os resultados esperados.

Se for criado um mecanismo que possibilite o fornecimento desses graus de qualidade para as ferramentas de consultas analíticas, torna-se viável apresentá-las junto com os resultados das consultas. Conseqüentemente, o usuário final do *Data Warehouse* (tomador de decisão) contará com um instrumento a mais para orientar suas análises, aumentando a confiabilidade do processo de tomada de decisão como um todo.

1.3 OBJETIVO

O objetivo deste trabalho é determinar as características de um ambiente computacional que, utilizado em conjunto com o *Data Warehouse*, forneça indicadores da qualidade dos dados para serem apresentados ao usuário final. O contexto no qual o ambiente de

DW considerado nesta dissertação está situado tem uma abrangência interorganizacional, o que o torna particularmente diferente dos DWs alimentados, exclusivamente, de dados provenientes da própria organização.

O ambiente proposto, denominado Ambiente de Suporte à Qualidade de Dados em *Data Warehouse* (AQUAWARE), faz uso da abordagem de serviços *Web* (TSALGATIDOU, PILIOURA, 2002) para disponibilizar um serviço que provê informações de qualidade para ferramentas que atuem como clientes do DW (ferramentas de consultas analíticas, por exemplo). Essa abordagem foi escolhida por estar alinhada às necessidades particulares de ambientes federados, de contexto interorganizacional. Para armazenar as informações sobre qualidade, são utilizados um Repositório de Metadados e um conjunto de Cubos de Qualidade (ELIAS, MARTINHO, 2003). O paradigma *Goal Question Metric* (BASILI et al. 1994) foi utilizado para modelar qualidade no repositório, bem como para orientar a definição de um conjunto de objetivos, questionamentos e métricas, utilizados para instanciar o repositório. Como parte integrante deste ambiente, existe ainda uma Biblioteca de Componentes, responsável por implementar as rotinas que obtêm, armazenam e tornam disponíveis as informações de qualidade.

Destaca-se que a solução aqui apresentada tem o propósito de atuar como infra-estrutura para o fornecimento de informações de qualidade para ferramentas cliente, não contemplando a apresentação das mesmas para o usuário final, o que ficará a cargo da ferramenta cliente que poderá fazê-lo de forma análoga aos dados das consultas analíticas.

1.4 ESTRUTURA

Essa dissertação está organizada da seguinte forma: no capítulo 2 é apresentada uma revisão de conceitos sobre qualidade de dados. O capítulo 3 tece considerações sobre o contexto interorganizacional e apresenta uma visão geral sobre a abordagem de *Data Warehouses* federados. O Ambiente de Suporte à Qualidade de Dados em *Data Warehouse* (AQUAWARE) é proposto no capítulo 4, bem como são descritos os elementos que o compõem. No capítulo 5 o paradigma *Goal Question Metric* (BASILI et al. 1994) é utilizado para orientar a proposição de objetivos, questionamentos e

métricas para a avaliação da qualidade de dados no AQUAWARE. O capítulo 6 apresenta um exemplo de aplicação, que mostra a utilização do AQUAWARE para a obtenção de informações de qualidade num DW federado de contexto interorganizacional. Por fim, no capítulo 7 são descritas as conclusões do trabalho, suas contribuições e limitações, além de possíveis trabalhos futuros.

CAPÍTULO 2

QUALIDADE DE DADOS

Na chamada era das organizações baseadas em conhecimento, ou era do capitalismo intelectual, a informação é um recurso organizacional crítico. Assim, adotar uma estratégia para garantir a qualidade dos dados é fundamental .

De acordo com o dicionário Aurélio, qualidade é uma "propriedade, atributo ou condição das coisas ou das pessoas, capaz de distingui-las das outras e de lhes determinar a natureza. Numa escala de valores, qualidade permite avaliar e, conseqüentemente, aprovar, aceitar ou recusar qualquer coisa." (FERREIRA, 1999, p.1675).

A norma ISO 8402 (1994) define a qualidade como a totalidade de características de uma entidade que lhe confere a capacidade de satisfazer às necessidades explícitas e implícitas.

Qualidade de dados é um conceito complexo, porque possui significados diversos para diferentes pessoas. Portanto, gerar uma medida de qualidade aceitável para todos não é trivial. Para isso é preciso definir as características da qualidade que interessam e, então, decidir como a avaliação da qualidade será feita através delas (PINHO, 2001).

Neste capítulo faz-se uma revisão das pesquisas realizadas na área de qualidade de dados. Inicialmente são apresentadas considerações acerca da qualidade no contexto de *software*. A seguir, é apresentado um estudo mais profundo sobre qualidade de dados, incluindo um levantamento dos critérios para a avaliação da qualidade de dados encontrados na literatura. Uma vez que o principal objeto de estudo da presente dissertação é a qualidade de dados voltada para o ambiente de *Data Warehouse*, foi realizado também, um estudo sobre propostas existentes para a qualidade de dados no DW, apresentado na seção 2.5.

2.1 QUALIDADE DE SOFTWARE

No contexto de *software*, algumas abordagens surgiram para definir as etapas necessárias para se avaliar a qualidade de um produto de *software* (ISO 9126 ,1991) (ROCHA,1983).

A norma ISO 9126 (1991) fornece seis características subdivididas em subcaracterísticas, que possibilitam avaliar a capacidade de um produto de *software* em satisfazer aos itens que foram especificados como necessários. Além disso, essas características possibilitam avaliar as necessidades implícitas que devem ser identificadas e definidas, em conformidade com a definição de qualidade feita na norma ISO 8402 (1994). A tabela 2.1 apresenta as características e subcaracterísticas que avaliam a qualidade de um *software*, segundo a norma ISO 9126 (1991).

CARACTERÍSTICAS	SUBCARACTERÍSTICAS							
Funcionalidade	Adequação, Acurácia, Interoperabilidade, Conformidade, Segurança de Acesso							
Confiabilidade	Maturidade, Tolerância a Falhas, "Recuperabilidade"							
Usabilidade	Inteligibilidade, Apreensibilidade, Operacionalidade							
Eficiência	Comportamento em relação ao tempo, Comportamento em relação aos recursos							
Manutenibilidade	"Analisabilidade", Capacidade para ser modificado, Estabilidade, Testabilidade							
Portabilidade	Adaptabilidade, Capacidade para ser instalado, Conformidade, Capacidade de substituição							

Tabela 2.1 - Características e Subcaracterísticas da Qualidade de Software (ISO 9126, 1991)

Numa definição de qualidade aplicada a *software*, Rocha (1983) define qualidade de *software* como sendo um conjunto de propriedades a serem satisfeitas, em determinado grau, de modo que o *software* satisfaça às necessidades de seus usuários.

No modelo Rocha (1983) para avaliação da qualidade de um produto de *software* são utilizados os conceitos de objetivos de qualidade, fatores de qualidade, critérios, processos de avaliação, medidas e medidas agregadas. Esse modelo é compatível com a norma ISO 9126 (1991).

Os objetivos de qualidade são as propriedades gerais que o produto deve possuir. Já os fatores de qualidade determinam a qualidade do produto sob o ponto de vista dos diferentes usuários, em diversos níveis. Esses fatores podem ser compostos por subfatores, pois um fator pode não definir completamente um objetivo. Os critérios definem os atributos básicos (primitivos), passíveis de serem avaliados. Isso não seria possível através dos objetivos e fatores de qualidade. Os processos de avaliação

determinam as métricas a serem usadas para expressar as medidas. Medidas indicam o grau de presença, no produto, de um determinado critério. São os valores resultantes da avaliação de um produto, segundo um critério específico. Medidas agregadas indicam o grau de presença de um determinado fator e resultam da agregação das medidas obtidas da avaliação segundo os critérios.

O modelo Rocha define uma taxonomia de atributos obtidos de forma *top-down*, que inclui desde os atributos de mais alto nível (objetivos e fatores) até atributos primitivos e independentes, passíveis de serem avaliados (critérios). A relação entre esses elementos estabelece que os objetivos de qualidade são atingidos por fatores de qualidade, que por sua vez podem ser compostos de subfatores avaliados por meio de critérios.

Esse método foi particularizado para sistemas de informação hospitalar e para o prontuário médico computadorizado (CARVALHO, 1997). Ele também foi utilizado para a definição de elementos da avaliação da qualidade de dados pela não conformidade (PINHO, 2001). Em ambos os casos, o modelo Rocha foi utilizado como base para a definição de objetivos, fatores, subfatores, critérios e processo de avaliação no contexto de qualidade de dados.

Outros trabalhos particularizaram esse método, considerando diferentes domínios de aplicação, tais como: *software* científico (PALERMO e ROCHA,1989) (BAHIA, 1992) (COMERLATO et al., 1994), financeiro (BELCHIOR, 1992), educacional (CAMPOS, 1994a) (CAMPOS 1994b), sistemas especialistas (OLIVEIRA, 1995) e sistemas de informação (BLASCHEK e ROCHA, 1995).

Além destas abordagens, destaca-se em especial o paradigma *Goal Question Metric* (BASILI et al., 1994), que foi concebido inicialmente para avaliar falhas em projetos da NASA (*National Aeronautics and Space Administration*) (BASILI, WEISS, 1984) (BASILI, ROMBACH, 1988) e, desde então, tem sido utilizado com sucesso, ao longo dos anos, para a avaliação da qualidade de *software* (OIVO, BASILI, 1992). Ultimamente, sua utilização começou a ser estendida também para avaliação da qualidade de dados (BOBROWSKI et al., 1999). Ele parte do princípio de que é necessário traçar objetivos no início do processo de avaliação, para que se possa realizar medições de forma significativa. Esta é uma abordagem do tipo *top-down* que provê

diretrizes para a definição de métricas sem o conhecimento prévio de medidas específicas. Por este motivo, esta abordagem foi utilizada para modelar a qualidade de dados na solução proposta nessa dissertação. Ela é descrita na próxima seção.

2.1.1 O PARADIGMA GOAL QUESTION METRIC

Segundo o GQM (*Goal Question Metric*), os objetivos não devem ser avaliados diretamente, mas através de questionamentos que precisam ser respondidos durante a avaliação da qualidade. Esses questionamentos também não são respondidos diretamente, mas através de métricas aplicadas ao produto ou processo em questão. Assim, técnicas e algoritmos específicos são utilizados para derivar respostas aos questionamentos a partir de métricas.

O GQM está modelado em três níveis (figura 2.1) : Conceitual (*Goal* / Objetivos), Operacional (*Question* / Questionamento) e Quantitativo (*Metric* / Metrica).

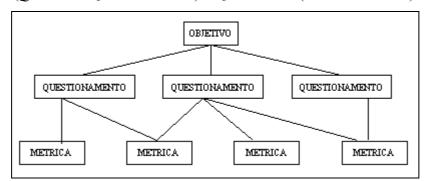


Figura 2.1 – Objetivos, Questionamentos e Métricas (BASILI et al., 1994)

Um objetivo é definido para um objeto, com diversos propósitos, em relação a vários critérios de qualidade, sob diferentes perspectivas, todos relativos a um determinado ambiente. Objetos mensuráveis podem ser:

- Produtos (Ex: especificações, projetos, programas, unidades de teste, dados)
- Processos (Ex: especificação, projeto, teste, entrevista)
- Recursos (Ex: pessoas, *hardware*, *software*, espaço físico de trabalho)

Além disso, objetivos pode ser refinados em subobjetivos.

Um conjunto de questionamentos é utilizado para caracterizar a forma como será realizada a avaliação de um objetivo específico. Os questionamentos buscam caracterizar o objeto mensurável em relação a critérios de qualidade, previamente selecionados, para determinar seu grau de qualidade sob a perspectiva considerada.

Métricas são conjuntos de dados associados a cada questionamento para que o mesmo possa ser respondido de forma quantitativa. A elaboração de métricas é um processo customizado, que varia conforme a situação e o contexto em questão.

O processo do GQM é composto pelos seguintes passos:

- Identificação de um conjunto de objetivos de qualidade no nível da corporação, divisão ou projeto;
- Derivação de questionamentos a partir dos objetivos;
- Especificação das medidas que precisam ser obtidas para responder os questionamentos; e
- Desenvolvimento de mecanismos para a obtenção das medidas.

2.2 TIPOS DE AVALIAÇÃO DA QUALIDADE DE DADOS EM BANCOS DE DADOS

Dois tipos de avaliações que podem ser feitas em um banco de dados, são: a avaliação quantitativa (ou objetiva) e a avaliação qualitativa (ou subjetiva) (PINHO, 2001).

Na avaliação quantitativa são utilizados indicadores objetivos para medir a qualidade da representação da estrutura do banco de dados. Neste caso, a medição da qualidade depende apenas do dado que está sendo medido, e não do ponto de vista do observador. São observados, por exemplo, valores do domínio, presença de valor, entre outros fatores.

A avaliação quantitativa da qualidade é normalmente realizada por ferramentas automatizadas (*software*). Esse tipo de avaliação está baseado na perspectiva do desenvolvimento tradicional, pois considera que a precisão e a objetividade são dois critérios necessários. Porém, a grande maioria dos usuários dos dados entende que eles não são suficientes para que os dados sejam considerados como de alta qualidade.

A avaliação qualitativa da qualidade depende do ponto de vista do observador e do dado que está sendo medido. Neste tipo de avaliação os indicadores subjetivos são as características da qualidade que foram definidas, objetivando capturar a percepção e a expectativa do usuário sobre a qualidade de dados. Ela pode ser realizada por especialistas ou usuários do banco de dados, ou ainda em conjunto com ferramentas automatizadas, que utilizam os indicadores subjetivos para julgar a qualidade atual do banco de dados e sua aptidão de uso.

2.3 ELEMENTOS PARA A DESCRIÇÃO DA QUALIDADE DE DADOS

Para avaliar qualidade é preciso antes conceituar e descrever o que se considera como dado de boa qualidade. Assim como a ISO 9126 (1991) utiliza características e subcaracterísticas para descrever qualidade de *software*, e o modelo Rocha (1983) utiliza objetivos, fatores, critérios, processos de avaliação, medidas e medidas agregadas com a mesma finalidade, as abordagens para a avaliação da qualidade de dados também utilizam elementos para descrevê-la. A maioria destas abordagens define uma estrutura hierárquica de conceitos para a caracterização da qualidade. Nesta dissertação serão utilizados os elementos definidos por Abate et al. (1998), que utiliza os conceitos de atributos, critérios (ou dimensões) e categorias da qualidade de dados.

Atributos da qualidade são características observáveis nos dados, possíveis de serem avaliadas e medidas. Os atributos podem ser agrupados em critérios (ou dimensões²) da qualidade. Os critérios da qualidade relacionam atributos que representam um conceito ou idéia consistente sobre a qualidade de dados, e que são geralmente avaliados em conjunto. Os critérios, por sua vez, podem ser agrupados em famílias, denominadas categorias. As categorias facilitam o reconhecimento de padrões na ocorrência de problemas relacionados com a qualidade.

Associadas aos critérios de avaliação da qualidade de dados, devem existir métricas que são o instrumento utilizado para a obtenção de medidas. As medidas indicam o grau de

² Na literatura o termo "dimensão" é geralmente utilizado para referenciar grupos de atributos da qualidade. Nesta dissertação o termo "critério da qualidade" é utilizado com esse significado, em lugar de "dimensão da qualidade".

presença dos critérios nos dados. Segundo Rocha (1983), medidas indicam o grau de presença, no produto, de um determinado critério. Alguns trabalhos encontrados na literatura (LEE et al., 2001) (PIPINO et al., 2002) (NAUMANN, 1998) (NAUMANN e LESER, 1999) (NAUMAN e ROLKER, 2000) (BOBROWSKI et al., 1999) representam iniciativas no sentido da definição de métricas para a mensuração da qualidade de dados. A utilização de métricas no processo de avaliação da qualidade de dados será discutida com mais detalhes no capítulo 5.

2.4 PROPOSTAS DE CRITÉRIOS PARA A AVALIAÇÃO DA QUALIDADE DE DADOS

Diversos estudos sobre a qualidade de dados já foram realizados, apresentando diferentes abordagens para a definição de critérios para a avaliação da qualidade. A qualidade de dados, conforme apresentada na literatura, é um conceito multidimensional. Acurácia, completeza, consistência e atualidade são critérios freqüentemente citados. Contudo, ainda não existe um consenso em relação a esses critérios. Numa tentativa de organização da literatura existente sobre qualidade de dados, Wang et al. (1995) utilizaram uma analogia entre qualidade de produtos e qualidade de dados, na criação de esquemas para a análise das pesquisas realizadas sobre a qualidade de dados.

Uma abordagem para a definição de critérios da qualidade, usando conceitos ontológicos, foi utilizada por Wand e Wang (1996). Essa abordagem baseia-se nos problemas que acontecem no mapeamento dos dados do mundo real para os sistemas de informação. Seu principal objetivo é fornecer orientações para os projetistas de sistemas, em questões relacionadas à qualidade de dados. Como resultado do trabalho realizado por Wand e Wang (1996), foi proposta a tabela 2.2, que contém os critérios para a qualidade de dados, relacionados às visões interna (ponto de vista técnico) e externa (ponto de vista dos usuários) dos sistemas de informação.

Visão	confiabilidade, "temporalidade", completeza, atualidade, consistência, precisão
Interna	precisao
Visão Externa	"temporalidade", relevância, conteúdo, importância, suficiência, facilidade de uso, utilidade, clareza, concisão, ser informativo, nível de detalhe, quantidade apropriada, escopo, facilidade de interpretação, facilidade de entendimento

Tabela 2.2 - Critérios da qualidade de dados nas visões interna e externa (WAND e WANG, 1996)

A dependência entre o contexto do usuário e a avaliação da qualidade de dados é defendida por Eckerson (2002). Segundo esse autor, o dado não precisa estar cem por cento livre de defeitos. Basta que ele atenda aos requisitos dos usuários. Os seguintes critérios são citados por Eckerson (2002) com o objetivo de caracterizar a qualidade dos dados, de uma forma geral: acurácia, integridade, consistência, completeza, validade dos domínios, disponibilidade e acessibilidade.

A questão da qualidade de dados, sob a perspectiva dos consumidores de dados (usuários finais), é também abordada por Wang et al. (1996). O objetivo principal de sua pesquisa foi elaborar um estudo empírico, voltado para a qualificação dos dados, que capturasse aspectos de qualidade importantes para os consumidores de dados e que culminasse com a elaboração de uma arquitetura hierárquica da qualidade de dados. Como resultado, esse trabalho propôs uma tabela (tabela 2.3) contendo os critérios agrupados em quatro categorias, na definição de dados de alta qualidade: intrínseca, de acessibilidade, de contexto e aspectos representacionais.

Categoria	Critérios									
QD Intrínseca	acurácia, objetividade, confiabilidade e reputação									
QD Acessibilidade	acessibilidade e segurança									
QD Contextual	relevância, valor agregado, "temporalidade", completeza e quantidade apropriada de dados									
QD Representacional	Facilidade de interpretação, facilidade de entendimento, consistência na representação e concisão									

Tabela 2.3 - Agrupamento de Critérios da Qualidade de Dados (WANG et al., 1996)

Os critérios e categorias definidos por Wang et al. (1996) foram utilizados em vários outros trabalhos sobre qualidade de dados (STRONG et al., 1997) (ABATE et al., 1998)

(PEREIRA JUNIOR, 2000) (PIPINO et al., 2002). A arquitetura proposta por Wang et al. (1996) foi utilizada no contexto de uma análise qualitativa, sob a perspectiva dos usuários dos dados, para identificar padrões de problemas de qualidade de dados (STRONG et al., 1997). Essa mesma arquitetura foi utilizada para definir requisitos de qualidade de dados, que podem ser aplicados no desenvolvimento e avaliação de Sistemas de Informação Ambientais (PEREIRA JUNIOR, 2000). Princípios para a definição de métricas para a qualidade de dados (PIPINO et al., 2002) também foram definidos com base nos critérios propostos por Wang et al. (1996). Esses mesmos critérios foram utilizados por Abate et al. (1998) para definir uma hierarquia para a análise de problemas relacionados à qualidade de dados.

Loshin (2001) analisou a qualidade de dados sob o ponto de vista da gestão do conhecimento. Esse autor define qualidade de dados como "adequação para o uso" e destaca que a avaliação da qualidade é dependente do contexto do usuário. Loshin baseou-se em Wang et al. (1996) para definir critérios para a qualidade de dados, que foram agrupados nas seguintes categorias: modelo de dados, valores dos dados, domínios da informação, apresentação dos dados e política da informação. A tabela 2.4 apresenta os critérios definidos por Loshin (2001), associados a cada uma das categorias citadas.

Modelo de Dados	clareza na definição, facilidade de compreensão, flexibilidade, robustez, conter apenas o essencial, granularidade dos atributos, precisão dos domínios, homogeneidade, naturalidade, capacidade de identificação, facilidade de obtenção, relevância, simplicidade, consistência semântica e consistência estrutural
Valores dos Dados	acurácia, tratamento de valores nulos, completeza, consistência e atualidade
Domínios da Informação	concordância dos usuários em relação à utilização, definição responsabilidade sobre a informação e ubiquidade
Apresentação dos Dados	adequação aos usuários, interpretação correta, flexibilidade, precisão do formato, portabilidade, consistência da representação, representação de valores nulos e espaço de armazenamento
Política da Informação	acessibilidade, definição de metadados, privacidade, redundância, segurança e custo unitário

Tabela 2.4 – Critérios e categorias da qualidade de dados, segundo Loshin (2001)

A definição de critérios de qualidade considerando o contexto interorganizacional foi abordada por Aubert et al. (2003) e Bertolazzi e Scannapieco (2001). Aubert et al.

definiram como critérios: acessibilidade, consistência, completeza, relevância, quantidade apropriada, "temporalidade", facilidade de interpretação e facilidade de entendimento. Os critérios propostos por Bertolazzi e Scannapieco (2001) são: acurácia, "temporalidade", completeza, atualidade, consistência, reputação, importância, acurácia sintática e acurácia semântica.

Um sumário de pesquisas realizadas no mercado e na área acadêmica, sobre a definição de critérios da qualidade de dados foi desenvolvido por Lee et al. (2001), o qual elaborou um quadro de referência que relaciona os critérios da qualidade aos autores que os propuseram. Dentre eles estão alguns autores já citados nesta seção (WANG et al.,1996) (WAND e WANG, 1996) e outros que também desenvolveram pesquisas sobre a definição de critérios da qualidade de dados (ZMUD, 1978) (JARKE e VASSILIOU, 1997) (DELONE e MCLEAN, 1992) (GOODHUE, 1995) (BALLOU e PAZER, 1998) (CYKANA et al., 1996) (MEYEN e WILLSHIRE, 1997) (KOVAK et al., 1997) (MANDKE e NAYAR, 1997) (MATSUMURA e SHOURABOURA,1996) (GARDYN, 1997) (REDMAN, 1992). Ao sumário criado por Lee et al. (2001) foram acrescentados os critérios propostos por Aubert et al. (2003), Bertolazzi e Scannapieco (2001), Loshin (2001) e Eckerson (2002). Esse quadro de referência é apresentado na tabela 2.5. O cabeçalho de cada coluna contém o número do autor a que se refere, de acordo com a legenda correspondente. O anexo II apresenta a descrição dos critérios.

Comparando os critérios da qualidade de dados apresentados na tabela 2.5 com os critérios para a qualidade no contexto de *software* (brevemente discutidos no capítulo 3), observa-se que alguns dos critérios da qualidade de dados recebem destaque também como critérios da qualidade de *software*. Exemplos são: acurácia, segurança de acesso e inteligibilidade (facilidade de entendimento) (ISO 9126, 1991).

Critério	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Total
Acurácia	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		17
Completeza	X		X	X		X	X	X	X		X		X	X	X	X	X	X	14
"Temporalidade"	X	X	X	X		X	X	X	X	X	X				X	X	X	X	14
Consistência			X			X	X	X			X			X	X	X	X	X	10
Confiabilidade	X	X	X	X	X		X		X	X	X								9
Facilidade de entendimento	X	X	X	X			X		X					X	X			X	9
Relevância	X		X	X			X		X					X	X			X	8
Acessibilidade	X		X		X				X			X	X			X		X	8
Atualidade				X	X		X						X	X	X		X		7
Existência de Metadados			X		X			X						X	X	X			6
Quantidade apropriada	X	X		X			X		X									X	6
Precisão		X		X			X							X	X				5
Importância				X			X							X	X		X		5
Representação consistente	X			X					X					X	X				5
Usabilidade	X			X	X		X		X										5
Detalhamento adequado					X		X							X	X				4
Unicidade				X				X						X	X				4
Segurança	X								X		X				X				4
Concisão	X			X				X	X										4
Utilidade			X	X			X												3
Privacidade			X								X				X				3
Reputação	X								X								X		3
Objetividade	X			X					X										3
Representação apropriada				X	X									X					3
Validade								X								X			2
Valor agregado	X								X										2
Naturalidade														X	X				2
Homogeneidade														X	X				2
Controle de redundância														X	X				2
Consistência estrutural														X	X				2
Portabilidade														X	X				2
Flexibilidade														X	X				2
Tratamento de valores nulos														X	X				2
Armazenamento eficiente														X	X				2
Facilidade de obtenção														X	X				2
Robustez														X	X				2
Ausência de ambigüidades					X		X												2
Registro de apelidos			X																1
Controle de versões			X																1

Critério	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Total
Origem identificada			X																1
Concordância sobre o uso															X				1
Integridade																X			1
Simplicidade															X				1
Acompanhamento															X				1
Ubiqüidade															X				1
Custo adequado															X				1

Legenda

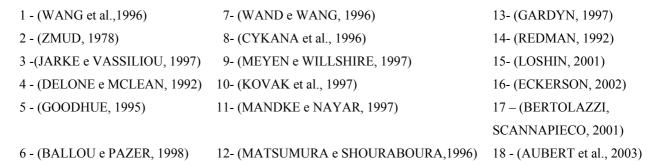


Tabela 2.5 – Quadro de Referência das propostas de critérios para a qualidade de dados

As iniciativas citadas até aqui surgiram no âmbito da Ciência da Computação. A seguir serão apresentadas propostas pertencentes à área da Ciência da Informação.

No contexto da Ciência da Informação, a avaliação da qualidade de dados pode ser feita sob diferentes aspectos. As duas grandes linhas de pensamento dominantes nessa área, bastante semelhantes à Ciência da Computação são: a vertente que enfatiza o produto (informação enquanto coisa) e a centrada no usuário (abordagem subjetiva). Nehmy e Paim (1998) selecionaram o modelo proposto por Marchand (1990) e as classificações de Repo (1989) e Olaisen (1990) como exemplares de tentativas de teorização da qualidade ou valor da informação. Essas classificações operam sem discriminar o tipo de abordagem em que se apóiam. As duas grandes linhas teóricas - "baseada no produto" e "baseada no usuário" - aparecem de modo combinado. A ênfase maior das categorizações está na identificação de atributos de qualidade, cada autor elegendo um determinado espectro deles, os quais, por diferentes vias de classificação, são agrupados sob os termos qualidade ou valor.

A desagregação do conceito de qualidade da informação em oito critérios interrelacionados é proposta por Marchand (1990). São eles: valor real, características suplementares, confiança, significado no tempo, relevância, validade, estética e valor percebido. Em relação ao critério valor real, Marchand faz alusão à variabilidade da percepção do valor do produto (informação ou serviço), dependente de estilos individuais de tomada de decisão. Sobre as características suplementares à utilidade básica de um produto ou serviço de informação, faz um alerta sobre os diferentes pesos que as características da informação podem ter em contextos diversos de tomadas de decisão. Com referência à confiabilidade, lembra a existência de atitudes contraditórias de confiança em relação a fontes. Sobre o critério significado no tempo, faz alusão à variabilidade da atualidade da informação em diferentes contextos de tomadas de decisão. Na definição de relevância, invoca as diferenças na percepção da relevância da informação entre projetistas de sistemas e agentes da tomada de decisão. Em relação ao critério validade, comenta sobre a variação da percepção da validade da informação, dependente de quem a fornece e de como é apresentada. Sobre a estética, menciona a subjetividade do aspecto estético da informação. E, finalmente, arrola a dimensão valor percebido, quando aponta a irracionalidade da atribuição de reputação pelo usuário a sistemas de informação.

Em artigo no qual se dispõe a demonstrar a contribuição de idéias e instrumentos de economistas para a análise prática do valor da informação, Repo (1989) chega à conclusão de que, embora haja interesse teórico sobre o tema "qualidade de dados", os estudos realizados são, de um modo geral, pouco produtivos em termos práticos, para se medir a informação. Além disso, acentua que, em geral, os estudos vindos da ciência da informação não fazem distinção entre valor de uso e de troca. Sugere, como um modo de se iniciar a organização da pesquisa, o uso da dupla abordagem do valor: valor de troca (caráter econômico) e valor de uso (caráter cognitivo). Diz Repo (1989, p.81) que "o aspecto filosófico (emocional,espiritual, social...)" conforma a base da análise do valor da informação. Entretanto, o próprio autor deixa em suspenso o valor filosófico ao afirmar que seu papel não será explorado no artigo, porque ele depende, na prática, dos indivíduos, e sua importância tende a refletir a avaliação de usuários individuais da informação. Os "valores práticos" dividem-se, para ele, em "valor de troca" de produtos da informação (serviço, canal ou sistema) e "valor de uso" que leva em consideração o

usuário, o uso e os efeitos do uso da informação, subdivididos em "valor esperado" e "valor percebido".

O caráter qualitativo da avaliação da informação no contexto da tecnologia eletrônica é abordado por Olaisen (1990). Partindo de um quadro de referência de corte fenomenológico, da teoria da administração de serviços e de um estudo empírico, esse autor propõe um modelo, onde aspectos usuais da qualidade (critérios) são agrupados em quatro categorias: qualidade cognitiva, qualidade do desenho da informação, fatores referentes ao produto da informação e fatores relativos à qualidade da transmissão. A qualidade cognitiva é dependente de como a fonte é valorizada pelo usuário. Nessa categoria, incluem-se os seguintes critérios: credibilidade, relevância, confiança, validade e significado no tempo . A qualidade do desenho da informação incorpora fatores referentes à forma: flexibilidade e seletividade. Os fatores referentes ao produto da informação são valor real e abrangência e, finalmente, os fatores relativos à qualidade da transmissão são definidos pelo critério da acessibilidade. O conjunto dos fatores e a relação entre eles configuram o que o autor denomina de processo de qualidade da informação.

Ao comparar os critérios da qualidade identificados nas abordagens acima, pertencentes à área de Ciência da Informação, com critérios para a qualidade de dados identificados no âmbito da Ciência da Computação (tabela 2.5), observa-se a existência de critérios da qualidade comuns a ambas as áreas, tais como: relevância, confiabilidade, validade, acessibilidade e reputação das fontes dos dados, entre outros.

Ainda no campo da Ciência da Informação, Campos (2001) discute a problemática representacional e do conhecimento e apresenta métodos e teorias para a modelização e representação do conhecimento, que podem ser utilizados para a organização do conceito de qualidade de dados.

2.5 ABORDAGENS PARA A QUALIDADE DE DADOS EM *DATA WAREHOUSE*

Nesta seção será apresentado um estudo sobre a qualidade de dados no ambiente de *Data Warehouse*. Foi realizada uma revisão de literatura, com o objetivo de identificar

que tipo de indicadores já foram propostos para avaliação da qualidade de dados no DW e em que etapas do processo eles são aplicados.

Inicialmente são apresentadas abordagens para a qualidade de dados segundo a visão de alguns autores como David Marco (MARCO, 2000a) (MARCO, 2000b), Jeff Rothenberg (ROTHENBERG, 1996), Inmon (INMON, 2003), Kimball (KIMBALL, 2000), Loshin (LOSHIN, 2001), Larry English (ENGLISH, 1999), o grupo de estudo do projeto *Data Warehouse Quality* (JARKE e VASSILIOU, 1997) e Vassiliadis (VASSILIADIS, 2000). A seguir, é feita uma comparação entre essas abordagens. Por último são traçados comentários sobre a qualidade de dados no contexto do *Commom Warehouse Metamodel* (OMG, 2003).

2.5.1 A ABORDAGEM DE DAVID MARCO

A abordagem de David Marco (MARCO, 2000a) (MARCO, 2000b) adota a utilização de metadados técnicos para armazenar informações sobre a qualidade dos dados do *Data Warehouse*. Eles são referenciados no nível de granularidade de uma linha no DW. Cada linha de dados é descrita a partir dos sistemas de origem, durante o processo de ETL (*Extraction, Transformation and Loading*) com o uso de metadados técnicos. Qualquer metadado associado à linha deve ser aplicável à linha inteira de dados, e não somente à maioria das colunas da tabela.

Abaixo, são apresentados exemplos de metadados propostos por David Marco (MARCO, 2000a) (MARCO, 2000b) que podem ser incorporados nos projetos de DW para facilitar medidas de qualidade de dados e promover a confiança do usuário na informação recebida.

• Data de Atualização: denota quando uma linha foi alterada pela última vez no *Data Warehouse*. Esse atributo mantém o significado histórico da informação no DW. É usado rotineiramente nos modelos dimensionais que implementam *slowly changing dimensions*³, para identificar quando a linha foi atualizada ou alterada.

_

³ Variação do conteúdo dos atributos das dimensões, ao longo do tempo (KIMBALL, 1998).

- Identificador do Ciclo de Armazenamento: identificador sequencial atribuído durante cada ciclo de armazenamento no *Data Warehouse*, independente da frequência de atualização (diariamente, semanalmente, etc.). Pode ser usado para remover facilmente os dados de um ciclo de armazenamento, se alguma questão de qualidade de dados surgir.
- Indicador Corrente: identifica a última versão de uma linha na tabela. Isso facilita
 a rápida identificação da última versão de uma linha nas comparações de data de
 realização.
- Identificador do Sistema Transacional: usado para guardar a origem de uma linha de dados no *Warehouse*. Esse metadado permite a identificação, para cada linha de uma tabela do DW, das origens usadas em sua construção, provendo aos usuários e desenvolvedores um poderoso meio para identificar e medir a qualidade do dado recebido de um sistema transacional. No caso onde o processo de ETL extrai e transforma o dado de muitos sistemas origem, é atribuído um valor ao metadado que representa essa integração particular.
- Indicador de Sistema Transacional Ativo: usado para indicar se as chaves de produção em uma tabela do *Data Warehouse* estão ainda ativas no sistema transacional original. Essa coluna pode ser usada efetivamente em uma variedade de atividades de análise para identificar dados não mais usados, ou dados que poderiam ser restringidos em relatórios.
- Indicador do Nível de Confiança: usado para indicar como as regras de negócio foram aplicadas durante os processos de ETL, para um registro de dados. Provê uma forma de medida, para o usuário, do nível de credibilidade da linha de dado, baseado no processo de transformação realizado. É usado para identificar problemas com a qualidade dos dados provenientes dos sistemas origem e para facilitar a correção dessas questões. Os níveis variam de 1 a 5, e vão desde dados de uma origem estável, até dados de origens externas, como novos serviços ou origens comerciais.

2.5.2 A ABORDAGEM DE JEFF ROTHENBERG

A abordagem apresentada por Jeff Rothenberg (ROTHENBERG, 1996) baseia-se em categorias de metadados requeridos para suportar a evolução e aperfeiçoamento da

qualidade de dados. Essas categorias são apresentadas em três níveis diferentes: banco de dados, elementos de dados e valores (instâncias) de dados. Segundo Rothenberg, a qualidade de dados não é um atributo binário, mas está diretamente relacionada ao contexto e ao propósito de uso. Uma abordagem abrangente de qualidade de dados requer a evolução da qualidade dos valores de dados, realizando VV&C (Verificação, Validação e Certificação) e a evolução dos processos que geram e modificam o dado, visando o aumento da qualidade dos dados que eles produzem.

A seguir são apresentadas as categorias de metadados propostas por Rothenberg (1996) para suportar a evolução e aperfeiçoamento da qualidade de dados. Essas categorias são tidas como necessárias, ou para evolução, ou para registro da qualidade de dados.

Metadados - nível de banco de dados

- Geral: descrição e significado do banco de dados (BD); uso ou variedade de propósitos e restrições do banco de dados; requisitos para acesso e uso; descrição e razão para estrutura ou projeto do banco de dados; relacionamentos globais com outros bancos de dados; informação do ciclo de atualização do BD.
- Informação de origem para o BD: origem e credibilidade da origem; informação de classificação, acessibilidade, capacidade de reprodução, autoridade de versão para o banco de dados.
- Caracterização: granularidade (nível de detalhe) e qualidade (exatidão, completeza,
 etc) pretendidas.
- Informação de elementos de dados: restrições e informação da dimensão de distribuição.
- Qualidade da medida (total e para cada uso): exatidão, consistência, completeza, clareza, flexibilidade, robustez do projeto de banco de dados, apropriação para o que se pretende usar.
- Informação para controle de processos

Metadados - nível de elementos de dados

- Significado do elemento de dado, seus meta-valores e metadados: definição do
 que o elemento de dados representa; significados de nulos (valor desconhecido,
 aplicabilidade de atributo, valores especiais, etc.); significados de exceções.
- Origem e informação do ciclo de atualização para o elemento de dado:
 permissão de múltiplas origens com ciclos de atualização múltiplos, irregulares;
 "modo de degradação" esperado, classificação, acessibilidade, possibilidade de
 reprodução, autoridade de versão.
- Informação de derivação/transformação: agregação ou outra informação de derivação; informação do processo de transformação; dados de controle do processo.
- Restrições, relacionamentos com outros dados/BDs: completeza de entidade e atributo, etc.
- **Domínios, tipos de dado e unidades de medida:** justificativa, portabilidade, flexibilidade, etc.
- Resolução, precisão, exatidão pretendida e esperada.
- Elemento de dado apropriado para o uso pretendido: significado, derivação, restrições, domínio, resolução, exatidão pretendida, etc.
- Histórico de mudanças: auditoria da evolução das escolhas de domínio, tipos e unidades; momento e origens das modificações dos elementos de dados.
- Auditoria VV&C: a respeito da adequação do elemento de dado, seu domínio, tipo, unidades, etc.

Metadados - nível de valores de dados

- Qualidade (total e para cada uso): exatidão, consistência (resultados de validação), datas de expiração, "modos de degradação", adequação para o que se pretende usar, origens e qualidade do metadado.
- Anotação: avisos, valores ou casos especiais, etc.
- Informação da origem: origem, derivação, tempo de geração e de entrada, etc.

- Informação da próxima origem: descrição de quando as alterações são esperadas e o que elas podem oferecer.
- Informação de derivação e transformação: agregação ou outra informação de derivação; informação do processo de transformação; dados de controle do processo.
- Auditoria de transformação: como esse valor tem sido transformado; informação das transações de transformação em progresso; momento e origens das modificações dos elementos de dados.
- Auditoria VV&C: finalidade da VV&C que tem sido feita nesse valor e "escopo" de validação e certificação.

2.5.3 A ABORDAGEM DE INMON

Inmon define a arquitetura da *Corporate Information Factory* (CIF), como uma evolução do ambiente de *Data Warehouse* (INMON, 2001a) (INMON, 2001b) (INMON, 2001c). A CIF detalha os diversos componentes que fazem parte de um ambiente de *Data Warehouse*, bem como a interação entre eles (figura 2.1). Nessa arquitetura, basicamente, os dados são provenientes do ambiente das aplicações transacionais, passam por uma camada de integração e transformação ao entrarem no ambiente da CIF, e são carregados no *Data Warehouse* Corporativo (EDW - *Enterprise Data Warehouse*) e/ou no Banco de Dados Operacional (ODS - *Operational Data Store*). O EDW é o coração da arquitetura da CIF. Ele contém os dados no menor nível de granularidade, de forma que os mesmos possam ser utilizados pelos demais componentes da CIF - *Data Marts* e *Exploration Warehouse*, por exemplo - formando uma única visão da realidade.

Segundo Inmon, o processo de garantia da qualidade de dados do DW, no ambiente da CIF pode ter três classificações, em função do local onde ocorrem (INMON, 2003). São elas:

- Na origem dos dados: sistemas legados, aplicações OLTP (On-line Transactional Processing), ERP (Enterprise Resource Planning), etc.;
- Durante a passagem dos dados para o DW, no processo de ETL;

• No próprio DW.

Existem vantagens e desvantagens associadas a cada uma das opções apresentadas acima, contudo Inmon sugere que as três sejam utilizadas de forma integrada, para uma maior eficácia. O principal objetivo da qualidade de dados nesse ambiente é garantir que o dado chegue até o DW com o mais alto grau de qualidade.

Contudo, uma coisa é assegurar que os dados que residem no DW são de alta qualidade. Outra, é dizer que os dados utilizados para análise também o são. Há uma divisão importante na CIF, que demarca as diferenças na abordagem para a qualidade de dados dentro desse ambiente. A figura 2.2 apresenta a linha que delimita duas áreas distintas, existentes na CIF, em relação à qualidade de dados. À esquerda está a área das aplicações, ETL e do próprio DW. Nesta área, o objetivo é limpar e corrigir os dados o máximo possível. À direita estão os *Data Marts* e o *Exploration Warehouse*. Nesta área pode ser necessário escolher que dados são melhores para serem usados na análise. No *Data Mart* e no *Exploration Warehouse*, a garantia da qualidade de dados implica em assegurar que os dados certos estão sendo usados para análise. A melhor garantia de que o dado mais adequado está sendo usado é obtida a partir da opinião do próprio usuário. Para isso, ele precisa saber qual a origem dos dados, o que eles significam e o quão atualizados eles estão. Essas informações são fornecidas através de metadados. Inmon (2003) não apresenta um detalhamento sobre esses metadados, sobre o local onde eles são armazenados, tampouco sobre como são apresentados ao usuário.

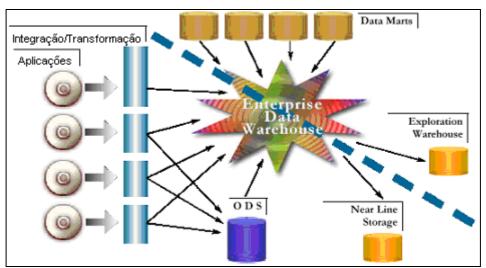


Figura 2.2 - Qualidade de Dados na CIF (INMON, 2003)

2.5.4 A ABORDAGEM DE KIMBALL

Kimball (2000) compara a qualidade de dados aos questionamentos feitos pelos usuários, ou mesmo por um auditor que examinasse o DW, sobre a origem, modo de obtenção, confiabilidade e completeza dos dados. Kimball (2000) apresenta uma lista de indicadores da qualidade que poderiam ajudar a responder a tais questionamentos. Eles estão divididos em três grupos: medidas quantitativas da qualidade de dados, indicadores da qualidade de dados no processamento e descritores do ambiente. A seguir é apresentada a lista desses descritores:

Medidas Quantitativas da Qualidade de Dados

- Pontuação sobre a avaliação geral da qualidade do dado;
- Completeza da coleta de dados comparada à melhor possível;
- Quantidade de elementos de dados implícitos;
- Quantidade de elementos de dados não aplicáveis;
- Quantidade de elementos de dados corrompidos;
- Quantidade de elementos de dados cujo valor está fora do intervalo definido pelo domínio;
- Quantidade de elementos cujo valor é desconhecido e que estejam sendo tratados como zero (ou outro valor padrão);
- Quantidade de elementos de dados alterados manualmente no processo de carga;
- Quantidade de elementos de dados não classificados em agregações regulares;
- Quantidade de correções realizadas após a carga dos dados.

Indicadores da Qualidade de Dados no Processamento

- Data/hora da conclusão da etapa de extração;
- Data/hora da geração da chave substituta (*surrogate*);
- Data/hora em que o dado foi examinado;

- Data/hora da criação/atualização da agregação;
- Data/hora da disponibilização on-line;
- Data/hora da última correção realizada.

Descritores do Ambiente

- Número da versão do sistema de ETL;
- Versão da lógica de alocação;
- Versão do planejamento;
- Versão do orçamento;
- Versão da região de vendas;
- Versão da conversão de moedas.

Segundo Kimball, o ideal é que essas informações sejam armazenadas no menor nível de granularidade possível. Uma abordagem possível é a associação direta dos indicadores de qualidade aos dados pela adição de uma chave de auditoria (chave substituta típica de DW) na tabela de fatos, correlacionando a mesma à dimensão que contém os indicadores. A figura 2.3 apresenta um exemplo de tabela de fatos no menor nível de granularidade, associada à dimensão de auditoria. A dimensão auditoria contém os indicadores para a qualidade de dados para essa tabela em particular. Os campos da dimensão auditoria possuem significados bem definidos para cada registro de fatos. A dimensão de auditoria provavelmente terá uma cardinalidade baixa quando comparada à tabela de fatos. As consultas realizadas com a dimensão auditoria serão semelhantes às realizadas com as demais dimensões.

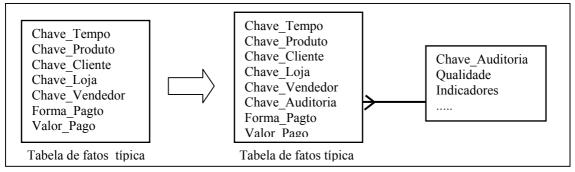


Figura 2.3 - Modelagem Dimensional da Auditoria da Qualidade de Dados

Quanto à carga dos dados de qualidade nas dimensões de auditoria, Kimball cita que a maior parte dos indicadores de qualidade é gerada e carregada no DW durante o processo de ETL.

2.5.5 A ABORDAGEM DE LOSHIN

Loshin (2001) define critérios para a qualidade de dados com base nos critérios propostos por Wang et al. (1996). Ele sugere a utilização desses critérios no processo de avaliação de conjuntos de dados (independente do local de armazenamento: DW, sistemas legados, etc) para a classificação dos requisitos de qualidade e para a definição de metas para os indicadores. Os critérios propostos por Loshin foram apresentados na seção 2.4.

Loshin (2001) entende que o conhecimento embutido em conjuntos de dados pode ser expresso através de regras de negócio e propõe uma arquitetura para a definição de regras para a qualidade de dados. Uma vez formalizadas as regras, é possível validar os dados frente a elas, para verificar se as informações utilizadas correspondem às expectativas dos usuários.

No contexto do *Data Warehouse*, Loshin adverte que muitos projetos falham por não dar a devida atenção aos requisitos de qualidade de dados e à avaliação da qualidade das informações que são importadas para o DW. A garantia de que os dados que entram no DW são de alta qualidade, diminui a possibilidade de erros nos processos de tomada de decisão. Loshin defende a idéia de que os dados devem ser validados antes de serem incluídos no *Data Warehouse*, ao invés de se fazer a limpeza após sua carga. Para isso, é necessário trabalhar em conjunto com os usuários dos dados, com o objetivo de identificar os requisitos de qualidade, determinar as regras de qualidade e integrar ferramentas para testar e validar os dados no momento da sua inserção no DW. Ao mesmo tempo, esse processo fornece subsídios para a identificação das causas dos problemas relacionados à qualidade.

Loshin exemplifica o processo de "certificação do *Data Warehouse*" com base na qualidade de dados e em regras de negócio, como uma forma de medir a confiabilidade dos dados armazenados no DW. As expectativas de qualidade são expressas através de

regras e a certificação é feita pela validação dos dados em relação a elas no momento de sua importação para o DW, conforme ilustrado na figura 2.4.

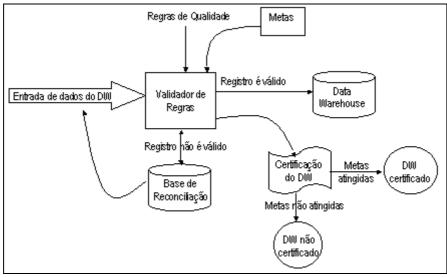


Figura 2.4 - Certificação do Data Warehouse (LOSHIN, 2001)

O primeiro passo do processo de certificação é a definição das regras de qualidade. A seguir, essas regras são importadas para um validador de regras. A cada regra é associado um valor referente à meta esperada para aquele requisito (os dados que apresentarem um resultado fora do intervalo definido na meta não atendem ao requisito). Conforme os registros passam pelo validador de regras, as regras relevantes para ele são testadas. Se nenhuma regra falhar, o registro é considerado válido e é importado com sucesso para o DW. Se alguma regra falhar, as informações sobre as regras violadas são anexadas a ele, e o mesmo é carregado numa base de reconciliação. Esses registros também podem ser importados para o DW, mas devem ser marcados como não estando em conformidade com as expectativas dos usuários. O total de falhas e sucessos é computado para cada regra. Ao final, a validade para cada regra é calculada pela divisão do total de registros carregados com sucesso pelo total geral de registros. Se os percentuais de validade das regras estiverem dentro das metas, o DW é certificado como atendendo aos requisitos do usuário. Caso contrário, ele não será certificado. Nesse caso, a base de reconciliação deve ser utilizada para identificar a causa dos problemas, de forma que eles possam ser corrigidos na origem.

2.5.6 A ABORDAGEM DE LARRY ENGLISH

De acordo com Larry English (1999), a qualidade da informação é composta por três fatores: definição clara ou significado dos dados, valores corretos e facilidade de entendimento do formato apresentado para o usuário final. A baixa qualidade em algum desses três componentes pode causar falhas nos processos de negócio e resultar em tomadas de decisão incorretas. Segundo Larry English, informação é a aplicação dos dados, e pode ser representada pela seguinte equação:

Informação = f (Dados + Definição + Apresentação)

Logo, a qualidade da informação requer que haja qualidade tanto na definição quanto no conteúdo dos dados. A qualidade na definição dos dados aplica-se a conceitos. Ela é uma característica e uma medida dos modelos de dados produzidos pelo processo de desenvolvimento das aplicações. A qualidade do conteúdo dos dados é o grau no qual os valores dos dados representam corretamente os fatos ou entidades do mundo real, e atendem às necessidades de informação dos clientes. Ela se aplica às instâncias dos dados.

Qualidade na Definição dos Dados

English (1999) compara definição dos dados e arquitetura da informação com especificação de produto, e define três aspectos da qualidade da informação como especificação de produto: qualidade na padronização dos dados, qualidade na definição dos dados e qualidade na arquitetura da informação. Características específicas da informação como produto incluem:

- Qualidade na padronização dos dados;
- Qualidade na definição dos dados:
 - Qualidade dos nomes dos dados: qualidade dos termos do negócio, clareza nos nomes das entidades, clareza nos nomes dos atributos, consistência dos tipos de domínio, qualidade nos acrônimos e abreviações dos termos do negócio, consistência dos nomes ou títulos dos dados.
 - Qualidade das definições dos dados: qualidade na definição dos termos do negócio, qualidade na definição das entidades, qualidade na definição dos

atributos, consistência entre o nome e a definição dos dados, valores do domínio corretos.

- Qualidade das regras de negócio: completeza e acurácia das regras de negócio.
- Qualidade na arquitetura da informação: relacionamentos corretos, clareza no modelo das informações de negócios, completeza e correção do modelo dos dados operacionais, completeza e correção do modelo do *Data Warehouse*, projeto do banco de dados para dar suporte aos processos operacionais, projeto e arquitetura dos bancos de dados distribuídos, projeto dos bancos de dados que dão suporte aos processos estratégicos e de tomada de decisão.

Qualidade no Conteúdo dos Dados

English (1999) define duas categorias básicas para as características qualidade da informação (conteúdo): a qualidade inerente e a qualidade pragmática. Qualidade inerente à informação é a medida da acurácia na representação dos fatos do mundo real pelos dados. Qualidade pragmática da informação é a medida do quão bem as informações habilitam os tomadores de decisão (clientes da informação) a atingirem os objetivos do negócio de forma eficiente e eficaz, e a satisfazer os clientes da organização.

Os critérios da qualidade inerente à informação, citados por English (1999) são: conformidade com a definição dos dados, completeza (dos valores), validade ou conformidade com as regras de negócio, acurácia em relação à origem, acurácia em relação à realidade, precisão, não duplicação, equivalência dos dados redundantes ou distribuídos, acessibilidade.

As características pragmáticas da qualidade da informação, citadas por English (1999) são: "temporalidade", clareza no contexto, integridade nas derivações, usabilidade, completeza ou corretude dos fatos.

Em sua abordagem, English (1999) não descreve onde os metadados sobre a qualidade devem ser armazenados, mas afirma que a qualidade dos dados não deve ser medida apenas dentro do *Data Warehouse*. Ela deve ser medida nos seguintes locais:

- Nos bancos de dados de origem dos registros, para verificar a eficácia do processo de criação dos dados;
- Nos bancos de dados referenciados pelos registros, para verificar a eficácia das atualizações dos dados e da respectiva propagação; e
- No Data Warehouse, para verificar a eficácia das transformações realizadas nos dados e da respectiva propagação.

2.5.7 A ABORDAGEM DO DWQ

O *Data Warehouse Quality* (DWQ) é um projeto cooperativo do programa *ESPRIT European Communities*, direcionado para o estabelecimento de bases para qualidade na administração do *Data Warehouse*, e que propõe a união de modelos semânticos da arquitetura do DW para explicitar modelos de qualidade de dados (JARKE e VASSILIOU, 1997).

De acordo com o DWQ, um sistema de qualidade de dados deve englobar a estrutura organizacional, com a definição de responsabilidades, procedimentos, processos e recursos para implementar a gerência desta qualidade. Esta gerência trata de todo o planejamento e de um conjunto de ações sistemáticas necessárias para a satisfação dos requisitos de qualidade.

Segundo a proposta do DWQ, como o DW é um sistema composto de diversos subsistemas e processos entre eles, deve ser introduzido um mapeamento entre os componentes do DW e o modelo de qualidade. A figura 2.5 relaciona parâmetros de qualidade aos aspectos de projeto e operação do DW.

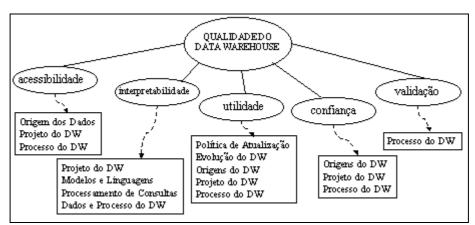


Figura 2.5 - Ligação entre fatores de qualidade e atividades do DW (JARKE e VASSILIOU, 1997)

O principal diferencial da proposta DWQ é a representação de questões de qualidade como consultas executáveis em um meta banco de dados semanticamente rico. A linguagem de meta modelagem utilizada foi a Telos, uma linguagem para representar o conhecimento em sistemas de informações (JEUSFELD et al.,1998).

Cada objeto, em qualquer nível e perspectiva da arquitetura do DW, pode ser medido sob o enfoque da qualidade, sendo que o modelo de qualidade é parte integrante do repositório de metadados. A informação de qualidade é relacionada aos objetos da arquitetura. Deste modo, tomadores decisão podem representar explicitamente seus objetivos de qualidade no repositório de metadados e, ao mesmo tempo, haverá um relacionamento entre a medida do objeto e os valores de qualidade correspondentes a este objeto.

No metamodelo do repositório, a qualidade é modelada segundo o paradigma GQM (BASILI et al., 1994), descrito na seção 2.1.1. Os requerimentos de alto nível do usuário são modelados como objetivos e as métricas de qualidade são valores que expressam alguma medida de propriedade do objeto. A figura 2.6 apresenta uma visão conceitual simplificada do metamodelo de qualidade do DWQ. O "DW object type" refere-se a qualquer meta-objeto da arquitetura do DW.

Uma meta de qualidade é um requisito abstrato, relacionado a um objeto, que se refere a um tomador de decisão, um critério e um propósito. Como tal, uma meta de qualidade é uma declaração em linguagem natural, formulada por um tomador de decisão, sobre o seu requisito de qualidade.

Um critério da qualidade é um termo usado para formular objetivos de qualidade. Os termos são organizados em uma hierarquia de especialização. Cada tomador de decisão pode ter suas próprias hierarquias preferidas. Exemplos são: qualidade de uso de dados, acessibilidade, consistência, completeza etc. A diferença entre critério da qualidade e objetivo da qualidade é que este último tem uma meta a alcançar e tem um objeto mensurável para o qual é aplicado. Critérios da qualidade provêem o vocabulário para formular objetivos da qualidade. Cada tomador de decisão pode ter um vocabulário diferente, i.e. uma hierarquia de critérios diferente.

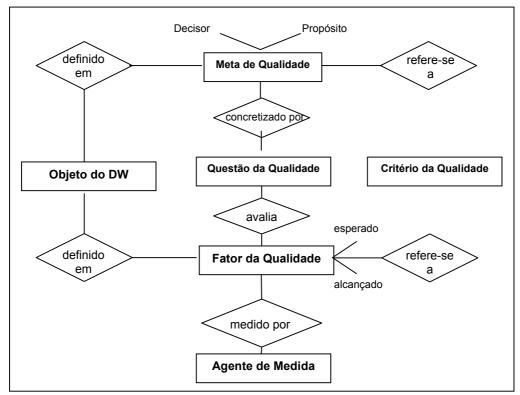


Figura 2.6 - Metamodelo de qualidade DWQ simplificado (JEUSFELD et al.,1998)

As principais vantagens do uso desta abordagem, através de um modelo semântico para gerência de qualidade no DW, são (JEUSFELD et al.,1998):

- Qualquer tipo de objeto mensurável pode ser representado no meta banco de dados do DW. Por exemplo, objetos estáticos, como definições de esquema, e objetos dinâmicos, como tradutores (*wrappers*), são suportados;
- Metas de qualidade podem ser formuladas a partir das perspectivas de um extenso conjunto de tomadores de decisão. Cada tomador de decisão pode avaliar a qualidade do DW a partir de sua perspectiva, validando as consultas de qualidade, ligadas às suas metas de qualidade;
- Consultas sobre a qualidade são consultas executáveis no meta banco de dados. As respostas delas são o termômetro para um tomador de decisão decidir se a qualidade é apropriada ou não. A qualquer momento, consultas sobre a qualidade podem ser inseridas, estendidas, modificadas ou removidas. Isto é devido ao fato que o meta banco de dados não ser somente um repositório de CASE (Computer-Aided Software Engineering) mas uma parte integrante do sistema de Data Warehouse; e

Medidas de qualidade são explicitamente armazenadas no meta banco de dados.
 Materializando-se sucessões de medidas de qualidade do mesmo tipo no meta banco de dados, pode-se observar tendências, através de consultas sobre a qualidade apropriadas.

A interdependência entre medidas de qualidade não foi abordada pelo DWQ. No entanto, cita-se que é possível assumir que a qualidade de alguns componentes do DW pode depender de outros. Por exemplo, a completeza de um cubo de dados materializado, que é um indicador de qualidade, depende da completeza das fontes de dados.

2.5.8 A ABORDAGEM DE VASSILIADIS

Segundo Vassiliadis (2000), o ponto de partida para a gerência da qualidade no DW é o repositório de metadados. Para isso, ele define uma arquitetura para o repositório, que modela o DW sob três perspectivas: arquitetural, de processos e da qualidade. O arquitetural abrange os componentes estáticos do *Data Warehouse* e o de processos, os componentes dinâmicos. A qualidade corresponde ao grau no qual são atendidas as expectativas das pessoas envolvidas com o ambiente de *Data Warehouse* em relação à mesma. Conseqüentemente, foi elaborado um metamodelo para cada perspectiva. O metamodelo arquitetural e o de processos não serão apresentados aqui, por estarem além do escopo desta dissertação. Vassiliadis (2000) apresenta uma descrição detalhada dos mesmos, bem como do metamodelo da qualidade, que está descrito a seguir.

O metamodelo da qualidade de Vassiliadis foi desenvolvido em paralelo com a proposta do DWQ (JARKE e VASSILIOU, 1997) e foi diretamente influenciado por esta. Conseqüentemente, esses metamodelos são bastante semelhantes. A abordagem de Vassiliadis (2000) para modelagem da qualidade foi igualmente concebida com base no paradigma GQM (BASILI et al., 1994) e também suporta a associação de medidas de qualidade qualquer meta-objeto da arquitetura do DW, sendo que a principal diferença entre esta abordagem e a do DWQ está nas entidades utilizadas para representar o GQM. O metamodelo de Vassiliadis (2000) pode ser instanciado no nível dos metametadados, no nível dos metadados ou no nível dos dados. Ele é parte integrante do repositório de metadados. Por apresentar todas essas características, o metamodelo de

Vassiliadis foi utilizado como fonte de inspiração para a elaboração do metamodelo do Repositório de Metadados da Qualidade (capítulo 4) da solução proposta nessa dissertação.

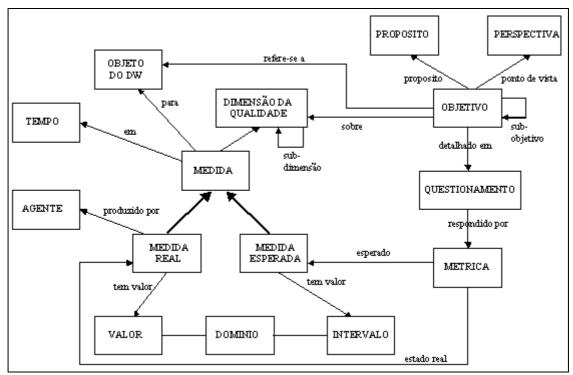


Figura 2.7 – O Metamodelo da Qualidade de Vassiliadis

A figura 2.7 apresenta o metamodelo da qualidade de Vassiliadis, que é composto pelas seguintes entidades:

- **OBJETIVO**: Finalidade que se pretende atingir com a avaliação da qualidade. Corresponde ao "*Goal*" do GQM (BASILI et al., 1994).
- **PROPOSITO**: ação que o OBJETIVO se propõe a realizar. Exemplos: avaliar, aprimorar, prever, controlar, etc.
- **PERSPECTIVA**: ponto de vista sob o qual OBJETIVO foi definido. Exemplos: usuário final, administrador do *Data Warehouse*, etc.
- DIMENSÃO DA QUALIDADE: critério de qualidade endereçado pelo OBJETIVO. Exemplo: confiabilidade, completeza, acurácia, etc.

- **OBJETO DO DW**: corresponde a qualquer objeto do ambiente de *Data Warehouse*, pertencente a qualquer nível da arquitetura (conceitual, lógica ou física) sobre o qual são aplicadas medidas de qualidade.
- **QUESTIONAMENTO**: esforço que caracteriza a forma pela qual o OBJETIVO será avaliado. Corresponde à "*Question*" do GQM (BASILI et al., 1994).
- METRICA: caracterização da ação de obtenção de um conjunto de informações sobre um OBJETO DO DW, que respondam ao questionamento quantitativamente.
- **MEDIDA**: valor resultante da medição do OBJETO DO DW em relação a uma determinada METRICA, num dado momento (TEMPO)
- MEDIDA REAL: tipo de medida que representa que foi realizada uma quantificação do grau de qualidade de um OBJETO DO DW por um AGENTE, produzindo um determinado VALOR.
- MEDIDA ESPERADA: tipo de medida que define o INTERVALO de valores permitidos para uma MEDIDA REAL. O INTERVALO deve ter o mesmo DOMINIO que os VALORES produzidos.
- VALOR: valor de uma MEDIDA específica.
- INTERVALO: faixa de VALORES (aceitáveis) para uma métrica.
- **DOMINIO**: domínio dos VALORES esperados e reais.
- **AGENTE**: artefato de *software* que produz as MEDIDAS REAIS.
- **TEMPO**: momento em que a MEDIAD foi obtida.

2.5.9 COMPARAÇÃO DAS ABORDAGENS APRESENTADAS

Todas as abordagens apresentadas nesta seção, possuem o propósito de implementar a qualidade de dados em *Data Warehouse*. Contudo, elas possuem algumas diferenças, principalmente com relação ao escopo abrangido por seus elementos, quanto às etapas do processo de ciclo de vida do DW a que se referem e quanto ao nível dos objetos sobre os quais é possível armazenar informações de qualidade. O nível refere-se à camada, na arquitetura de modelagem multi-camadas, ao qual o objeto pertence. A

arquitetura de modelagem multi-camadas é decrita de forma detalhada no capítulo 4 (seção 4.1). Por ora, destaca-se que são quatro as camadas dessa arquitetura e que elas referem-se aos seguintes níveis: instâncias dos dados, modelos (ou esquemas), metamodelos, e meta-metamodelos. A tabela 2.6 apresenta uma comparação entre essas abordagens.

Autor	Escopo	Nível dos objetos	Etapas
David Marco	Informações sobre a qualidade no	Instâncias dos dados	ETL
(MARCO, 2000a)	nível de linha, nas instâncias de dados.		
(MARCO, 2000b)			
Jeff Rothenberg	, ,	Modelos	Todo o ciclo de
(ROTHENBERG,	níveis de banco de dados, elementos	Instâncias dos dados	vida do DW.
1996)	de dados e instâncias de dados.	3.75. 1 . 11	
Inmon	Não detalha	Não detalha	Ambiente OLTP
(INMON, 2003)			ETL
TZ: 1 11	1.0 ~ 1.1.1	T (A : 1 1 1	DW
Kimball	Informações sobre a qualidade no	Instancias dos dados	ETL
(KIMBALL, 2000)	menor nível de granularidade, ligado		
т 1'	à instância do dado, na tabela de fatos	T (A : 1 1 1	ETI
Loshin	Validação das instâncias dos dados	Instancias dos dados	ETL
(LOSHIN, 2001)	frente a regras de qualidade.	M - 1-1	DDi
Larry English	Metadados sobre a qualidade da		DB origem
(ENGLISH, 1999)	,,	Instâncias dos dados	DB referenciado DW
DWO	apresentação dos dados. Define um metamodelo para a	Metamodelos	Todo o ciclo o
DWQ (JARKE e	Define um metamodelo para a qualidade, onde objetivos de qualidade		de vida do DW.
VASSILIOU, 1997)	(relacionados às dimensões da	Instâncias dos dados	de vida do Dw.
VASSILIOU, 1997)	qualidade) podem ser associados a	ilistalicias dos dados	
	objetos do DW.		
Vassiliadis	3	Metamodelos	Todo o ciclo de
(VASSILIADIS, 2000)	qualidade, onde objetivos de qualidade		vida do DW.
(11351111115, 2000)	(relacionados às dimensões da	Instâncias dos dados	vidu do D W.
	qualidade) podem ser associados a		
	objetos do DW.		

Tabela 2.6 - Comparação entra as abordagens para a qualidade de dados no DW

Ainda em relação ao nível dos objetos que podem ser referenciados em cada proposta, destaca-se que as abordagens de Rothenberg (1996) e English (1999) apenas definem metadados para objetos no nível dos modelos, mas não especificam como esses objetos seriam referenciados. Da mesma forma, as abordagens do DWQ (JARKE e VASSILIOU,1997) e de Vassiliadis (2000) possibilitam o armazenamento de informações de qualidade sobre objetos pertencentes a qualquer nível, mas nenhuma delas detalha como seria a implementado o mecanismo para referenciar esses objetos na prática.

Todas essas abordagens são ainda semelhantes, no sentido que consideram o ambiente de *Data Warehouse* no contexto de uma organização. Nelas, os dados que alimentam o DW são provenientes de sistemas e fontes internos à organização. Esse ambiente difere daqueles em que o DW é composto também de dados externos, provenientes de outras organizações. Esse é um cenário bastante comum em órgãos da administração pública, como agências reguladoras, que, internamente, geram produtos como resultado da expertise e competência aplicadas aos dados coletados externamente e que são divulgados sob diferentes formas de produtos informacionais; e centros de processamentos de dados voltados para informações da administração pública.

Observa-se atualmente uma necessidade crescente de soluções para a qualidade de dados em projetos que vão além das fronteiras da organização, envolvendo informações provenientes de fontes externas. Considerações sobre *Data Warehouses* em contexto interorganizacional são apresentadas no capítulo 3.

2.6 QUALIDADE DE DADOS E O COMMON WAREHOUSE METAMODEL

O *Common Warehouse Metamodel* (CWM) (OMG, 2003) é um padrão criado pela OMG (*Objetct Management Group*) para a representação dos metadados do ambiente de *Data Warehouse*. O principal propósito do CWM é proporcionar uma fácil integração de metadados de sistemas de DW, sistemas de inteligência de negócio, ferramentas de DW e repositórios de metadados em ambientes heterogêneos distribuídos.

O metamodelo do CWM está dividido em cinco pacotes que contêm os submetamodelos que representam os metadados do ambiente de *Data Warehouse*. São eles:

• **Object Model**: subconjunto da UML (*Unified Modeling Language*) que provê elementos básicos para a criação e descrição de metamodelos. É nesse pacote que está a classe ELEMENT, que é a classe de mais alto nível do CWM, a qual exerce o papel de superclasse para qualquer outra classe do metamodelo, uma vez que todas as outras estão direta ou indiretamente subordinadas ela;

- Foundation: coleção de pacotes de metamodelos que contêm elementos do modelo, que representam conceitos e estruturas compartilhados pelos demais pacotes do CWM;
- Resource: contém os metamodelos que representam recursos de dados nos formatos: orientado a objetos, relacional, multidimensional e XML (eXtensible Markup Language);
- Analysis: inclui os metamodelos que representam as transformações de dados,
 OLAP (On-line Analytical Processing), mineração de dados, visualização da informação e nomenclatura de negócio;
- Management: contém os metamodelos que representam os processos e o resultado das operações do ambiente de Warehouse.

O metamodelo do CWM não define estruturas específicas para modelar qualidade de dados no DW, contudo ele pode ser estendido para representar conceitos e estruturas relacionados à qualidade. Ele contempla a possibilidade de mensuração de características de objetos do modelo, através da classe MEASUREMENT (figura 2.8), pertencente ao pacote *Warehouse Operation* (OMG, 2003).

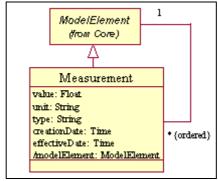


Figura 2.8 – A classe MEASUREMENT do CWM (OMG, 2003)

A classe MEASUREMENT está associada a MODELELEMENT (classe diretamente subordinada a ELEMENT), que é uma superclasse em relação às demais classes mensuráveis do CWM. Essa associação possibilita que seja realizada uma mensuração sobre qualquer outra classe do CWM (pois todas são subclasses de MODELELEMENT). Contudo, o escopo de MEASUREMENT no CWM (figura 2.8) está limitado ao nível dos metamodelos, da arquitetura de modelagem multi-camadas, não contemplando a mensuração de objetos em todos os níveis.

CAPÍTULO 3

DATA WAREHOUSES FEDERADOS E O CONTEXTO INTERORGANIZACIONAL

Numa economia global, descentralizada e direcionada à informação, onde empresas competem e cooperam em escala mundial, as necessidades de integração, troca e compartilhamento de informações têm-se tornado cada vez mais intensas. Nesse cenário, as decisões tomadas no dia-a-dia das empresas não mais baseiam-se apenas em informações internas, mas também em informações provenientes de fontes externas como fornecedores, parceiros, clientes e outras instituições com as quais compartilhe objetivos de negócio. Sendo assim, cada vez mais as tecnologias de apoio ao processo de tomada de decisão, como o *Data Warehouse*, passam a integrar um contexto bem mais abrangente e começam a trabalhar com dados interorganizacionais.

A abordagem de ambientes federados pode ser considerada como uma entre as diversas soluções tecnológicas para promover a integração de informações provenientes de organizações autônomas e com objetivos comuns.

Neste capítulo serão feitas considerações sobre *Data Warehouses* situados em contextos interorganizacionais e a sua relação com a abordagem de *Data Warehouses* federados, bem como são apresentadas duas propostas de arquiteturas para ambientes de DW federados. Dando continuidade, serão apresentadas algumas soluções correntes para a interoperabilidade em ambientes OLAP.

3.1 DATA WAREHOUSE EM CONTEXTO INTERORGANIZACIONAL

A grande motivação para o surgimento da tecnologia de *Data Warehouse* foi o volume crescente de dados disponíveis às organizações, que trouxe como consequência a maior necessidade de integração das informações. A função de um ambiente de *Data Warehouse* é tornar as informações corporativas, obtidas a partir de bancos de dados operacionais e de fontes de dados externas à organização, acessíveis para entendimento e uso das áreas estratégicas de uma organização. O crescimento da utilização do DW como a principal ferramenta de apoio às tomadas de decisões motivou as empresas a realizarem análises mais ricas e abrangentes, incentivando-as a integrar seu DW com

dados provenientes de organizações externas, como clientes, fornecedores, parceiros e outras. Neste caso, o DW assume uma abrangência interorganizacional, onde as informações internas são combinadas com informações externas, produzindo informações interorganizacionais.

Laudon (1986) examina a qualidade de dados sob a ótica de um sistema interorganizacional. Algumas dificuldades apontadas por essse autor são: a falta de uma autoridade centralizadora ou gerencial, pouca padronização quanto ao conteúdo e formato da informação e pouca precisão. A avaliação da qualidade dos dados em DWs situados em contexto interorganizacional é mais importante e complexa do que em DWs alimentados apenas com informações internas, porque muitas vezes, a qualidade dos dados não é conhecida e não existem critérios que possam guiar uma avaliação, ainda que preliminar, da vertente qualidade e integridade dos dados. Alguns dados podem ser fornecidos informalmente ou podem não estar embasados em instrumentos regulatórios vigentes. A qualidade pode apresentar disparidades entre as diversas partes das bases de dados. Alguns fatores que aumentam a complexidade do processo de gerência sobre a qualidade dos dados nesse contexto são a heterogeneidade dos ambientes tecnológicos, diferenças entre as estruturas dos dados das organizações, diferenças na granularidade dos dados, diferenças na sintaxe e na semântica dos dados, e diferenças nos critérios que cada instituição utiliza na avaliação da qualidade dos dados que produz.

Na maioria dos casos, as informações interorganizacionais estão espalhadas em diversos *Data Marts* (ou DWs) internos ou externos à organização. Provavelmente esses *Data Marts* (ou DWs) possuem diferentes esquemas de representação e estruturas de descrição, mesmo quando tratam do mesmo domínio de negócio. Isto torna bastante complexo o problema da integração e interoperabilidade entre eles. A abordagem de *Data Warehouses* federados provê uma arquitetura propícia à integração e interoperabilidade entre *Data Marts* e/ou *Data Warehouses* heterogêneos, autônomos e distribuídos, tendo sido objeto de pesquisas nos últimos anos. A seção 3.2 apresenta um breve resumo sobre este tópico.

3.2 DATA WAREHOUSES FEDERADOS

A crescente demanda por integração de dados para subsidiar, com mais precisão, as tomadas de decisão foi um fator motivador para a evolução das federações de dados, que provêem uma coleção de bases de dados independentes, visualizados logicamente como uma única fonte de dados. O conceito de federação é utilizado também no âmbito de *Data Warehouse*. O *Data Warehouse* federado provê uma interface que o torna semelhante um grande DW, mas que na realidade, apenas adiciona camadas sobre os DWs existentes, possibilitando a execução de consultas sobre um grande *Data Warehouse* "virtual" (HUBER et al., 2001). Ambientes federados favorecem a integração de informações em ambientes heterogêneos, autônomos e distribuídos, característicos de situações que envolvem diversas organizações autônomas, com objetivos comuns.

Diversas são as propostas de arquiteturas para ambientes de *Data Warehouse* federados (ANAND,1997) (MANGISENGI,HUBER,2001). Em grande parte das abordagens, essa arquitetura é estruturada em diversos níveis, sendo que a principal diferença entre elas é em relação à materialização ou não do elemento situado na camada de mais alto nível, que corresponde ao esquema federado. A seguir são apresentadas duas abordagens para a arquitetura do *Data Warehouse* federado. São elas: a abordagem de Anand (1997), que propõe a materialização do esquema federado e a abordagem de Mangisengi e Huber (2001), onde o referido esquema não está materializado.

Devido à independência de esquema e plataforma entre os DWs que compõem o ambiente federado, fica evidenciada a necessidade de uma tecnologia de integração, que possibilite o acesso a dados multidimensionais independente de plataforma. Algumas soluções para interoperabilidade em ambiente OLAP são apresentadas na seção 3.3.

3.2.1 A ARQUITETURA PROPOSTA POR ANAND

Anand (1997) propõe um modelo integrado para o *Data Warehouse* federado que define uma arquitetura de três níveis: no nível mais baixo estão os bancos de dados operacionais e no nível mais alto, o *Data Warehouse*. O nível intermediário é composto por uma interface que provê os métodos de acesso entre os dois outros níveis. Nesta

arquitetura, ilustrada na figura 3.1, o *Data Warehouse* é modelado como uma visão materializada do esquema federado.

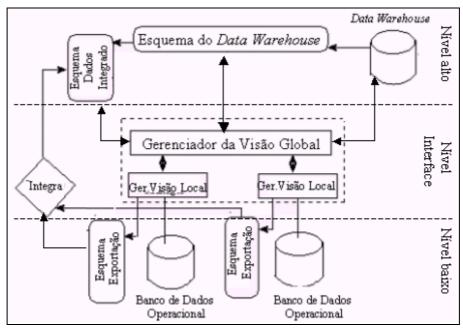


Figura 3.1 – A Arquitetura de DW Federado proposta por Anand (1997)

Quando uma consulta é executada, os dados solicitados são recuperados de forma transparente pelos bancos de dados operacionais, através da camada de interface. Os dados requisitados são inicialmente representados de maneira uniforme através dos esquemas de exportação. Os esquemas de exportação, depois de integrados e resolvida a heterogeneidade semântica entre eles, são combinados com os dados externos e mapeados para o esquema integrado de dados pelo Gerenciador de Visão Global, que também é responsável por mapear o esquema integrado para o esquema do *Data Warehouse*. Além disso, o Gerenciador de Visão Global é responsável por coordenar os diversos Gerenciadores de Visão Local, que mapeiam os esquemas externos para os esquemas locais dos bancos operacionais.

3.2.2 A ARQUITETURA PROPOSTA POR MANGISENGI e HUBER

A arquitetura de Mangisengi e Huber (2001) adota a abordagem de múltiplos níveis combinada ao conceito de mediadores, provendo maior flexibilidade e modularidade. Ela é apresentada na figura 3.2.

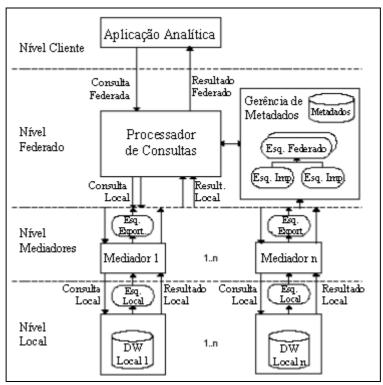


Figura 3.2 – A Arquitetura de DW Federado proposta por Mangisengi e Ruber (2001)

O nível local é composto por DWs distribuídos, autônomos e heterogêneos, que são descritos por esquemas locais. O nível dos mediadores funciona como uma interface entre o nível local e o nível federado. Ele é responsável por transferir, formatar e traduzir os dados dos DWs locais para o nível federado e vice-versa. O mediador recebe consultas do nível federado e as traduz para a linguagem de consulta dos DWs locais. Além disso, ele traduz o esquema local para o esquema de exportação, que segue o modelo de dados canônico do esquema federado, mantendo apenas as partes do esquema local que serão exportadas para o esquema federado. No nível federado o esquema de exportação é denominado esquema de importação. Este nível provê uma visão unificada dos diferentes DWs, através do esquema federado, criando uma visão homogênea dos mesmos. Ele é composto por um processador de consultas e por um gerenciador de metadados. O processador de consultas recebe a consulta federada do nível cliente, decompõe a mesma em diversas subconsultas, utilizando as informações do repositório de metadados, e as distribui para os mediadores. Quando os DW locais retornam os resultados, o processador de consultas faz uma consolidação dos mesmos e envia o resultado final para o nível cliente. Finalmente, o nível cliente provê consultas federadas para os usuários do DW federado. A principal diferença dessa proposta em relação à abordagem de Anand (1997) é que neste caso, os dados permanecem armazenados apenas nos diversos DWs locais, sendo que o DW federado consiste apenas de metadados.

3.2.3 CONSIDERAÇÕES SOBRE O ESQUEMA FEDERADO

Conforme visto nas duas últimas seções, a principal diferença entre as arquiteturas de Anand (1997) e Mangisengi e Huber (2001) diz respeito à materialização ou não do esquema federado. As principais vantagens da materialização são a maior velocidade na execução das consultas (principalmente quando esta envolve informações detalhadas, provenientes de diferentes DWs locais) e a menor probalilidade de ocorrerem inconsistências, que poderiam ser eliminadas na etapa de ETL do DW de nível mais alto. Como as consultas federadas são executadas diretamente nesse DW, a carga de processamento é menor, reduzindo os requisitos de hardware e, conseqüentemente, os gastos com equipamentos. Em contrapartida, a não materialização do esquema diminui o intervalo de tempo necessário para tornar os dados disponíveis para consulta, porque elimina a etapa de ETL do DW de nível mais alto. Além disso, é mais simples adicionar um DW local à federação, pois não é necessário executar o referido ETL novamente, sempre que isso acontecer.

A princípio, a materialização é uma boa escolha quando a velocidade na execução das consultas for um aspecto crítico e quando houver restrições em relação a gastos com equipamentos. Além disso, a janela de tempo disponível para carga de dados deve ser suficiente para comportar o processo de ETL do DW de nível mais alto. Quando esta não for suficiente, será melhor optar pela abordagem não materializada, que também é uma boa escolha quando a volatilidade dos dados é muito alta.

O que se pode concluir é que não existe uma solução rígida, única e definitiva. A escolha da alternativa mais adequada irá variar em função do contexto atual da organização, do ambiente tecnológico, da equipe técnica e de negócios.

3.3 INTEROPERABILIDADE EM AMBIENTE OLAP

Nesta seção são apresentadas algumas soluções para a interoperabilidade em ambiente OLAP. De forma geral, essas tecnologias apóiam-se em padrões abertos, como XML e Java, e utilizam a *Web* como infra-estrutura de comunicação. A seção 3.3.1 apresenta a

solução de Mangisengi e Huber (2001), que utiliza XML para prover interoperabilidade no ambiente de DWs federados; a seção 3.3.2 apresenta o XML for Analysis (MICROSOFT, HYPERION, 2002) e a seção 3.3.3, a Java OLAP Interface (HYPERION, SUN, 2003).

3.3.1 INTEROPERABILIDADE NA ARQUITETURA DE MANGISENGI e HUBER

Mangisengi e Huber (2001) apresentam uma proposta que utiliza a XML para prover interoperabilidade no ambiente de *Data Warehouses* federados, apresentado na seção 3.2.2. Nela, documentos XML e XML DTDs (*Document Type Definition*) são utilizados para representar os esquemas e modelos de dados, conforme ilustrado na figura 3.3.

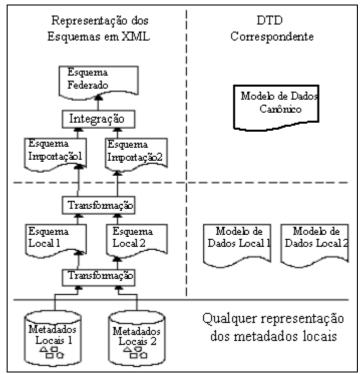


Figura 3.3 – Interoperabilidade em DWs Federados usando XML (MANGISENGI, HUBER, 2001)

A figura 3.3 está dividida em duas colunas. A coluna da esquerda contém a representação dos esquemas federado, exportação, importação e locais. A coluna da direita contém os DTDs relacionados às representações dos modelos de dados em XML (canônico e locais). Uma vantagem da utilização da XML para a interoperabilidade é que os arquivos nesse formato são independentes de plataforma.

3.3.2 XML FOR ANALYSIS

O XML for Analysis (XMLA) tem como proposta prover interoperabilidade em ambientes OLAP heterogêneos, utilizando a *Web* como infra-estrutura de comunicação (MICROSOFT, HYPERION, 2002).

O XMLA é um padrão comercial aberto para interface de serviço *Web*, projetado especificamente para processamento OLAP e para funções de mineração de dados (MICROSOFT, HYPERION, 2002). Ele tem como objetivo fornecer uma API (*Application Program Interface*) de acesso padrão aberta para os provedores de OLAP do mercado. Ele define um modo padrão para conexão a fontes de dados OLAP ou servidores de banco de dados multidimensionais, análogo ao ODBC (*Open DataBase Connection*) para os bancos de dados relacionais. Além disso, prevê consultas a dados e metadados, usando a sintaxe de uma linguagem de consulta padrão.

Essa linguagem é o mdXML, que é uma versão encapsulada da linguagem MDX (*Multidimensional Expressions*). A MDX é uma linguagem de expressão multidimensional definida na especificação OLE DB para OLAP (MICROSOFT, 2003). A versão atual da especificação do mdXML provê apenas o elemento <Statement></Statement>, que pode conter uma sentença MDX para os provedores multidimensionais (MICROSOFT, HYPERION, 2002).

A arquitetura do XMLA, ilustrada na figura 3.4, é construída sobre os padrões de serviços *Web* (TSALGATIDOU, PILIOURA, 2002) já existentes como HTTP (*Hypertext Transfer Proptocol*), XML e SOAP. Ela fornece a base para interoperabilidade entre tecnologias de inteligência de negócio fracamente acopladas, provenientes de vários vendedores e executando numa variedade de *hardwares* e sistemas operacionais através da *Web* e em ambientes privados distribuídos. O serviço *Web* do XMLA consiste de uma interface baseada em padrões (HTTP, XML, SOAP), utilizada para a acessar componentes não padronizados. Ele recebe mensagens SOAP, extrai a sentença mdXML contida na mensagem SOAP e passa-a para o mecanismo de Serviço Analítico para processamento através da API Java. O resultado é passado de volta para o serviço *Web* XMLA através da API Java, onde é envelopado em uma mensagem SOAP e enviado para o cliente que requisitou a consulta.

O XMLA possui uma interface simplificada, com apenas 2 métodos: *Discover* e *Execute*, além de informações de conexão a cada chamada método (*properties*). O método *Discover* é utilizado para recuperar metadados, sendo que as informações retornadas dependem dos valores dos parâmetros passados como entrada. O método *Execute* é utilizado para requisitar uma ação ao servidor, que poderá ser uma consulta ou uma atualização de dados. O anexo I contém a especificação dos métodos *Discover* e *Execute* do XMLA.

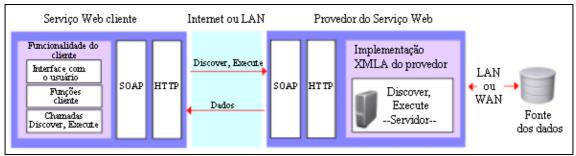


Figura 3.4 – Exemplo de arquitetura usando XMLA

3.3.3 JAVA OLAP INTERFACE

A Java OLAP *Interface* (JOLAP) é uma API orientada a objetos, totalmente desenvolvida na linguagem Java, projetada para o ambiente J2EE (Java 2 *Enterprise Edition*), que suporta a criação e manutenção de dados e metadados OLAP de uma maneira independente de vendedor (HYPERION, SUN, 2003). Ela oferece um conjunto de objetos, classes e métodos padrão para aplicações de inteligência de negócio. A JOLAP não possui uma linguagem de consulta multidimensional definida em sua especificação. Ao invés disso, possui vários métodos, classes e modelos de consultas disponíveis (HYPERION, SUN, 2003).

A especificação JOLAP é uma extensão do trabalho feito na JMI (*Java Metadata Interface*) para o CWM, que oferece definições comuns de estruturas de dados, objetos, aplicações e regras de negócios em um DW. Ela oferece uma definição completa da Interface Java OLAP: sua arquitetura conceitual, modelo lógico, interfaces programáticas Java, integração com J2EE, e suporte a padrões abertos para integração de sistemas e interoperabilidade.

No modelo JOLAP (HYPERION, SUN, 2003), apresentado na figura 3.5, as classes JOLAP estão organizadas em pacotes interdependentes.

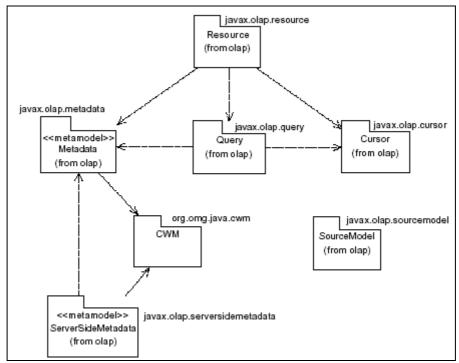


Figura 3.5 – O modelo JOLAP (HYPERION, SUN, 2003)

O pacote Resource contém as classes que representam as conexões a serviços JOLAP.

O pacote *Metadata* baseia-se no metamodelo OLAP do CWM e contém as classes que modelam os objetos do mundo OLAP. Este pacote replica as classes do CWM OLAP, necessárias para que os clientes JOLAP possam formular consultas usando o modelo *Query*.

O pacote *ServerSideMetadata* define classes de metadados adicionais, que correspondem às classes do CWM OLAP voltadas para a parte de implantação.

O pacote *Query* define todas as classes que compõem o modelo JOLAP *Query* de alto nível. Conceitos fundamentais acerca de consultas OLAP como dimensões, seleções, cubos, filtros, ordenações, membros calculados, etc. são representados nesse modelo.

O pacote *Cursor* contém as classes que representam a navegação e gerência dos resultados produzidos a partir da avaliação de instâncias do modelo *Query*.

Finalmente, o modelo *Source* define um conjunto de interfaces adicionais que podem ser utilizadas para implementar ou estender o modelo *Query*.

A JOLAP está alinhada com os metamodelos Multidimensional e OLAP do CWM, o que significa que ela suporta a construção e publicação de aplicações de *Data*

Warehouse e inteligência de negócio, ferramentas e plataformas baseadas nos padrões abertos da OMG, para especificações de metadados e sistemas: MOF (*Meta Objects Facility*), XMI (*XML Metadata Interface*), UML e CWM.

3.3.4 COMPARAÇÃO DAS PROPOSTAS DE INTEROPERABILIDADE

Nesta seção é feita uma comparação entre as propostas de interoperabiliade em ambiente OLAP. A proposta de Mangisengi e Huber (2001) considera apenas a interoperabilidade no nível dos esquemas e modelos de dados, tendo uma abrangência bem mais restrita que as outras duas. Por este motivo a comparação aqui apresentada tem como foco apenas os padrões XMLA e JOLAP.

JOLAP e XMLA são esforços para definir um mecanismo padrão de troca de dados e metadados para a área de inteligência de negócio. Eles atendem à necessidade de oferecer protocolos e plataformas comuns de comunicação e programação para quebrar as barreiras para a troca efetiva de informação entre aplicações que fazem gestão de negócios, servidores OLAP, DW, ferramentas de consulta e relatórios, entre outros. Apesar de ambos terem como objetivo prover interoperabilidade entre ambientes analíticos heterogêneos, eles não competem entre si, podendo ser utilizados de forma complementar. Uma solução baseada em XMLA pode fazer uso do JOLAP para fazer a comunicação com bancos de dados multidimensionais. O JOLAP, por sua vez, oferece suporte à arquitetura de serviços *Web* baseados em HTTP, XML e SOAP (EVERETT, 2003).

Uma vantagem importante do JOLAP é a conformidade com os padrões da OMG, o que ainda não está explícito na especificação XMLA. Uma vantagem relevante do XMLA é o fato de possuir uma linguagem padrão para consultas multidimensionais, o que é não é trivial de se fazer no JOLAP, pois é realizado através de linhas de comandos que utilizam vários métodos e classes. A tabela 3.1 apresenta uma comparação destes dois padrões.

Características		XMLA	JOLAP
Ambiente		Serviços Web	J2EE
Linguagem de consulta		mdXML	-
Padrões suportados		SOAP, WSDL	CWM, MOF, UML, XMI
Especificação		Incompleta – novembro de 2002	Completa – agosto de 2003
Linguagem de		Todas (é independente de linguagem)	Java
programação suportada			
Escopo		interações analíticas	DW como um todo
Objetivos		Fornecer uma interface padrão para troca	Fornecer uma interface
		de informações analíticas	padrão para troca de
			informações analíticas
Perspectiva futura de		. linguagem de consulta padrão para	API comercial padrão para
acordo com Hyperion		inteligência de negócio	desenvolvedores Java e
		. padrão de interoperabilidade entre as	aplicações baseadas na
		ferramentas de inteligência de negócio	arquitetura J2EE
Principais	Em comum	Hyperion, SAS	Hyperion, SAS
empresas	Outras	Microsoft, BO, DSS, MicroStrategy,	Oracle, Sun, IBM, Unisys,
envolvidas		Simba, Brio, Cognos, Cristal Decisions	Nokia

Tabela 3.1 - Comparação XMLA x JOLAP

Em relação aos fabricantes, a Hyperion entende que o JOLAP e o XMLA são padrões complementares, ao invés de competitivos. Apesar de ser possível implementar XMLA sem o uso de JOLAP, a especificação JOLAP suporta a arquitetura de serviços *Web* que depende de servidores de aplicação, XML e mensagens SOAP. A implementação XMLA da Hyperion usa a API Java desenvolvida com base no trabalho de especificação do JOLAP, para comunicar com o servidor OLAP (servidor analítico Essbase XTD) (SPOFFORD, 2002).

A discussão em torno da competitividade entre as tecnologias, pode estar baseada no fato da sobreposição e similaridade das funcionalidades, e nos benefícios dos padrões. Além disso, a Microsoft está apoiando apenas o XMLA e a Oracle apenas o JOLAP.

Outro ponto de vista a ser analisado é a natureza das empresas que apóiam as tecnologias. As empresas que apóiam o XMLA são predominantemente empresas especializadas no desenvolvimento de ferramentas de inteligência de negócio, enquanto as que apóiam o JOLAP são empresas de desenvolvimento de vários tipos de soluções. Com isso o XMLA pode vir a se tornar um padrão para interoperabilidade entre as ferramentas de BI.

Tendo como base o cenário atual de ambas tecnologias, acredita-se que maiores pesquisas e a eminente evolução dos padrões subsidiarão futuras e mais precisas análises.

CAPÍTULO 4

AMBIENTE DE SUPORTE À QUALIDADE DE DADOS EM *DATA*WAREHOUSE – AQUAWARE

As decisões tomadas em uma organização podem desencadear ações diretamente relacionadas ao rumo da empresa, que poderão conduzi-la ao sucesso, mas que também poderão ser responsáveis pelo seu fracasso. As informações utilizadas como subsídios durante a tomada de decisão têm uma grande influência sobre o resultado das ações tomadas. Daí a grande preocupação das empresas em apoiar-se em informações de boa qualidade.

Neste capítulo serão apresentadas as características de um ambiente proposto para ser utilizado como infra-estrutura para o fornecimento de informações sobre a qualidade de objetos do ambiente de *Data Warehouse*, pertencentes a qualquer nível da arquitetura de modelagem multi-camadas, com o objetivo de apoiar os processos de tomada de decisão, denominado "Ambiente de Suporte à Qualidade de Dados em *Data Warehouse*" (AQUAWARE).

4.1 ARQUITETURA DE MODELAGEM MULTI-CAMADAS

A especificação de uma arquitetura de modelagem com múltiplas camadas, para definir e relacionar informações em diferentes níveis de abstração, não é uma preocupação nova. Na década de 80, um grupo de especialistas da ISO (*International Standardization Organization*) propôs uma arquitetura de referência para descritores, com o objetivo de estabelecer padrões para um Sistema de Dicionário de Recursos de Informação, denominado IRDS (*Information Resouce Dictionary System*) (OLLE, 1988). O IRDS define uma arquitetura de modelagem de quatro níveis (figura 4.1), descritos a seguir:

- Nível de aplicação: contém dados relevantes para o usuário do sistema de informação.
- Nível IRD: contém a definição dos dados no nível da aplicação.

- Nível de definição IRD: identifica os tipos de informação que podem ser armazenadas no nível IRD.
- Nível fundamental: contém as definições dos conceitos utilizados no nível de definição do IRD. Consiste dos tipos dos dados, cujas instâncias são registradas no nível de definição do IRD.

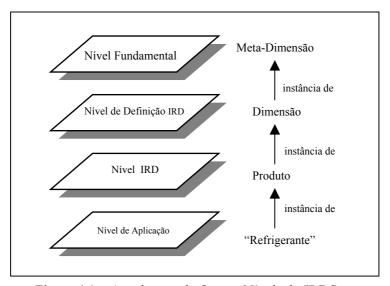


Figura 4.1 – Arquitetura de Quatro Níveis do IRDS

Posteriormente foi definida pelo OMG, uma arquitetura de metadados de quatro níveis, análoga à arquitetura de quatro níveis do IRDS, descrita a seguir:

- A primeira camada (M0) compreende as instâncias dos dados utilizados nas aplicações. Ela é conhecida como camada dos dados.
- A segunda camada (M1) possui os dados que descrevem as informações da camada situada logo abaixo. Ela é conhecida como camada dos metadados e refere-se aos modelos (ou esquemas).
- A terceira camada (M2) contém as descrições dos metadados. Ela é conhecida como camada dos meta-metadados e refere-se aos metamodelos. O CWM (OMG, 2003) encontra-se nessa camada.
- A quarta camada (M3) contém as descrições dos meta-metadados. Ela é conhecida como camada dos meta-meta-metadados e contém o MOF que é o meta-metamodelo capaz de descrever qualquer metamodelo.

Com o objetivo de padronizar a terminologia utilizada é apresentada abaixo uma analogia entre as abordagens do IRDS e da OMG, ao lado dos termos que serão utilizados ao longo desse trabalho, para referenciar os diferentes níveis da arquitetura de modelagem multi-camadas (tabela 4.1).

Termo	IRDS	OMG
Nível 0	Nível de aplicação	Camada dos dados ou M0
Nível 1	Nível IRD	Camada dos metadados ou M1
Nível 2	Nível de definição IRD	Camada dos meta-metadados ou M2
Nível 3	Nível fundamental	Camada dos meta-meta-metadados ou M3

Tabela 4.1 – Padronização da terminologia para os níveis da arquitetura multi-camadas

4.2. IDENTIFICAÇÃO DE OBJETOS NA ARQUITETURA DE MODELAGEM MULTI-CAMADAS

Para os casos em que é necessário se ter uma visão integrada e uniforme dos objetos da arquitetura de modelagem multi-camadas, independente do nível ao qual eles pertençam, deve haver um mecanismo capaz de identificá-los unicamente. A utilização de identificadores globais para referenciar objetos em qualquer nível de abstração, foi proposta por Campos (1993) para o contexto de bancos de dados relacionais, com o objetivo de integrar esquema e dados. Campos (1993) define um mecanismo de indexação global única para todos os objetos do banco de dados relacional, pertencentes a qualquer nível da arquitetura de modelagem multi-camadas, implementado através de identificadores gerados internamente pelo sistema (*surrogate*). Assim, a cada objeto do banco (dado ou metadado) é associado um identificador único, através do qual ele pode ser referenciado. Para relacionar objetos de diferentes níveis, Campos associou a cada objeto, uma referência para o objeto do qual ele é instância. A figura 4.2 ilustra essa solução.

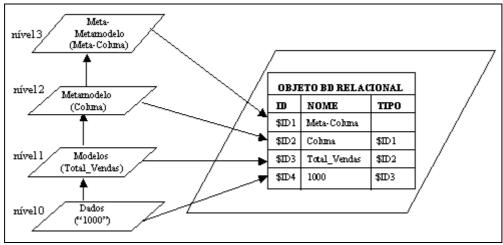


Figura 4.2 – Esquema simplificado da solução proposta por Campos (1993)

À esquerda na figura 4.2 estão alguns objetos do banco relacional, posicionados em seus respectivos níveis na arquitetura de modelagem multi-camadas. Conforme ilustrado à direita da figura, a utilização do identificador global único (ID) possibilita que dados e metadados sejam referenciados de maneira uniforme e integrada. O tipo (TIPO) relaciona cada objeto ao seu correspondente, no nível imediatamente acima da arquitetura multi-camadas. Assim, "Meta-Coluna", "Coluna", "Total_vendas" e "1000" são todos objetos do BD relacional, aos quais foram associados identificadores únicos (ID). Observando-se a informação do tipo (TIPO) é possível identificar o relacionamento entre esses objetos: "1000" tem o tipo \$ID3, logo é instância de Total_Vendas; "Total_Vendas" tem o tipo \$ID2, logo é instância de Coluna; "Coluna" (\$ID2) tem o tipo \$ID1, logo é instância de Meta-Coluna.

Apesar de ter o foco em bancos de dados relacionais, o mecanismo de indexação global única, definido por Campos (1993), consiste numa abordagem genérica, que pode ser reutilizada em outros contextos.

4.3 CARACTERÍSTICAS DO AMBIENTE DE SUPORTE À QUALIDADE DE DADOS EM *DATA WAREHOUSE* (AQUAWARE)

O AQUAWARE tem como objetivo possibilitar o fornecimento para um cliente, de informações sobre a qualidade de objetos do ambiente de *Data Warehouse*, pertencentes a qualquer nível da arquitetura de modelagem multi-camadas. O cliente pode ser outro DW, uma ferramenta OLAP, um processo de Extração/Transformação e Carga ou qualquer outro processo, ferramenta, banco de dados ou atividade que receba objetos do

DW como entrada e utilize-os de alguma forma. Além disso, o cliente pode estar dentro ou fora da organização.

Uma característica importante do ambiente aqui proposto é a possibilidade de associar informações de qualidade a objetos pertencentes a qualquer nível da arquitetura de modelagem multi-camadas. Isto porque a qualidade das informações apresentadas como resultado de uma consulta analítica está relacionada não apenas à qualidade das instâncias dos dados, mas também a objetos pertencentes aos demais níveis, como por exemplo:

- objetos pertencentes aos modelos (ou esquemas): uma dimensão ou uma variável de fato específicas, etc.;
- objetos relacionados aos processos do Warehouse: rotinas de carga, processos de transformação de dados, etc.;
- objetos pertencentes aos metamodelos: meta-objeto dimensão, meta-objeto variável de fato, etc; e outros.

O AQUAWARE é composto por um Repositório de Metadados da Qualidade (RMQ), um conjunto de Cubos de Qualidade, uma Biblioteca de Componentes e um Serviço de Qualidade que, ao serem incorporados ao ambiente de *Data Warehouse*, possibilitam o fornecimento de informações sobre a qualidade dos dados armazenados no mesmo. A figura 4.3 ilustra a arquitetura do AQUAWARE. Destaca-se que o Serviço Consulta não é parte integrante da solução aqui apresentada, representando apenas o mecanismo de execução de consultas no DW.

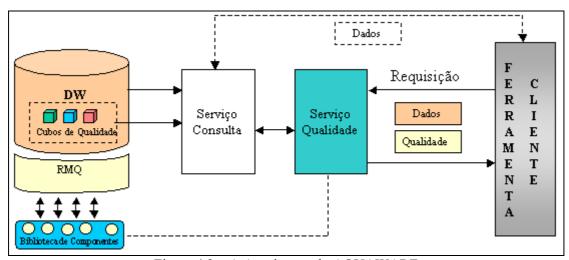


Figura 4.3 – A Arquitetura do AQUAWARE

No AQUAWARE, a avaliação da qualidade é realizada segundo o paradigma GQM (BASILI et al. 1994), apresentado na seção 2.1.1. Os parâmetros de avaliação são expressos em termos de objetivos, questionamentos e métricas, os quais ficam armazenados no Repositório de Metadados da Qualidade. Quando um cliente solicita informações sobre a qualidade de um objeto, ele o faz em termos dos objetivos. Cada objetivo é formado por um propósito, em relação a um critério de qualidade, para um objeto do DW, sob uma determinada perspectiva. Assim, um cliente pode, por exemplo, ter como objetivo avaliar (propósito) a confiabilidade (critério de qualidade) da variável de fato Total vendas (objeto do DW) sob a perspectiva do usuário final (perspectiva). A resposta retornada expressa o grau no qual o objetivo está sendo atingido, o qual é obtido a partir das medidas de qualidade. As medidas de qualidade correspondem às respostas aos questionamentos através das métricas. Elas são geradas por componentes presentes na Biblioteca de Componentes (a partir das métricas associadas aos objetos do DW) que as armazenam no RMQ. O mecanismo de indexação global única de Campos (1993), descrito na seção 4.2, foi utilizado para referenciar os objetos do DW no RMQ, para que os mesmos pudessem ser associados a medidas de qualidade. Assim, pressupõe-se que cada objeto tem uma identificação única no DW, que é utilizada para referenciá-lo no RMQ.

O Metamodelo do RMQ foi criado como uma extensão do CWM, onde o Metamodelo da Qualidade no DW de Vassiliadis (VASSILIADIS, 2000), descrito na seção 2.5.8, foi utilizado para representar as informações sobre a qualidade dos objetos do DW, segundo o paradigma GQM (BASILI et al., 1994). A definição do Metamodelo do RMQ como uma extensão do CWM possibilita a integração do AQUAWARE com qualquer ferramenta (banco de dados, OLAP, mineração de dados, etc.), cujo repositório de metadados esteja modelado segundo esse padrão (figura 4.4). Nesse caso, o RMQ passa a ser parte integrante do repositório de metadados da ferramenta. A seção 4.4 apresenta a descrição detalhada do Metamodelo do RMQ.

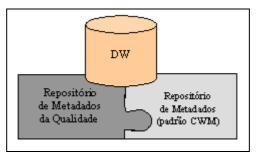


Figura 4.4 – Integração entre o Metamodelo do RMQ e o CWM

Após a geração das medidas de qualidade, outros componentes da Biblioteca são utilizados para calcular o grau no qual os objetivos estão sendo atingidos, para cada objeto do DW. Essas informações são armazenadas nos Cubos de Qualidade, que constituem a fonte de consulta sobre a qualidade dos dados do DW.

O Serviço de Qualidade é o responsável por disponibilizar as informações de qualidade para os clientes. Ele funciona como um mediador entre o cliente e o Serviço de Consulta. Ao receber uma consulta do cliente, ele verifica se o mesmo está solicitando dados, dados de qualidade, metadados dos cubos de dados ou metadados dos cubos de qualidade. Se foi feita uma requisição de dados ou de metadados dos cubos de dados, o Serviço de Qualidade repassa a consulta para o Serviço de Consulta, recebe o resultado e redireciona-o para o cliente. Se foram requisitados dados de qualidade, ele monta uma consulta aos Cubos de Qualidade, com base nos objetos do DW envolvidos na consulta original, e a envia para o Serviço de Consulta. O resultado é redirecionado para o cliente. Caso tenham sido solicitados metadados dos cubos de qualidade, o Serviço de Qualidade envia para o cliente a estrutura desses cubos. A consulta aos metadados dos Cubos de Qualidade tem como objetivo informar ao cliente que tipos de informação de qualidade estão disponíveis e qual a estrutura dessas informações (para que ele possa interpretar o resultado das consultas). Esse deve ser o primeiro tipo de consulta executada pelo cliente, que precisa fazê-lo apenas uma vez, pois após sua execução ele estará de posse dos tipos de informação de qualidade que poderá consultar e será capaz de interpretar os resultados recebidos. A especificação da forma de apresentação das informações de qualidade para o usuário final ficará a cargo da ferramenta cliente.

Assim, o Serviço de Qualidade funciona como um "cliente" do Serviço de Consulta, tornando o acesso aos Cubos de Qualidade transparente para a ferramenta cliente. Através dele, a ferramenta cliente pode requisitar informações de qualidade sem montar

consultas aos Cubos de Qualidade, bastando apenas fornecer a consulta aos cubos dos dados e informar que tipo de informações de qualidade pretende receber. Apesar do Serviço de Qualidade exercer o papel de mediador, nada impede que o cliente envie uma consulta diretamente para o Serviço de Consulta. O caso de uso da figura 4.5 ilustra a interação entre o cliente, o Serviço de Consulta e o Serviço de Qualidade.

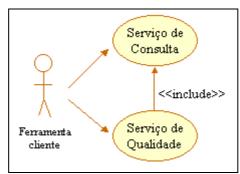


Figura 4.5 - Interação entre o cliente, o Serviço de Consulta e o Serviço de Qualidade

Devido características interorganizacional, a inerentes ao contexto heterogeneidade, complexidade e necessidade de padronização, o ideal é que exista um mecanismo e uma interface padrão para a solicitação e recebimento de informações sobre a qualidade. Por este motivo, foi utilizada a abordagem de componentes, disponibilizados como serviços Web, na implementação dos Serviços de Consulta e Qualidade. Na solução aqui proposta, a qualidade de um objeto do DW é obtida através de solicitações ao Serviço Web de Qualidade, que é implementado por componentes presentes na Biblioteca de Componentes. A interface padrão provida pelo serviço facilita a utilização dos componentes por qualquer tipo de cliente (seja ele um processo ETL, uma ferramenta OLAP, um DW, uma ferramenta de mineração de dados, etc.), esteja ele dentro ou fora da organização. A descrição dos componentes da Biblioteca é apresentada na seção 4.7 e a especificação desses componentes, no anexo IV.

4.4 O AQUAWARE EM *DW* FEDERADO INTERORGANIZACIONAL

No capítulo 2 foram tecidas considerações sobre *Data Warehouses* de abrangência interorganizacional e sobre o papel da arquitetura de *Data Warehouses* federados nesse contexto. Neste tipo de ambiente é muito importante que haja uma gerência sobre a qualidade dos dados e que exista um mecanismo que possibilite a apresentação de informações sobre a qualidade para o usuário final, pois como grande parte das fontes

de dados são externas, em geral, a geração e manutenção dos dados são feitas de forma autônoma e independente, e o usuário não tem como prever o quanto pode confiar nos resultados apresentados. A utilização do AQUAWARE representa um passo em direção à maior qualidade das tomadas de decisão baseadas em *Data Warehouses* interorganizacionais, onde é freqüente a utilização de arquiteturas federadas. No entanto, mesmo no contexto de uma única organização, sua utilização também pode ser interessante se as fontes de dados apresentam uma certa autonomia. Um exemplo é o caso de lojas que possuem várias filiais.

A arquitetura de DW federado de Mangisengi e Huber (2001), apresentada na seção 3.2.2, foi alterada para incorporar os elementos do AQUAWARE. Ela está ilustrada na figura 4.6. Apesar dessa arquitetura adotar a não materialização do esquema federado, nada impede que o AQUAWARE seja utilizado em soluções que materializem esse esquema, como na abordagem de Anand (1997), pois ele é flexível o suficiente para funcionar com quaisquer das duas abordagens. No capítulo 6 é apresentado um exemplo de aplicação, onde o AQUAWARE é incorporado a um ambiente cujo esquema federado está materializado.

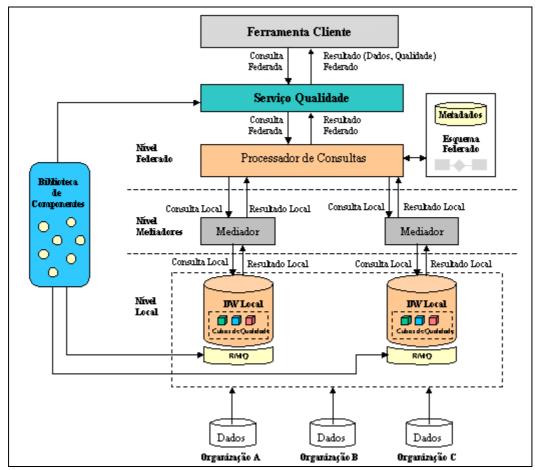


Figura 4.6 – O AQUAWARE em DW federado de contexto interorganizacional

Na arquitetura da figura 4.6, dados provenientes de DWs pertencentes a diferentes organizações são utilizados para compor um DW federado. Os *Data Warehouses* locais possuem um repositório de metadados com informações sobre a qualidade de seus dados. O metamodelo (esquema) desse repositório é comum para todos os DWs da federação, pois é utilizado o padrão definido para o AQUAWARE, apresentado na seção 4.5.3. A modelagem dos Cubos de Qualidade também é a mesma tanto para os esquema locais, quanto para o esquema federado e corresponde à definida para ao AQUAWARE, apresentada na seção 4.6. Seguindo a abordagem de Mangisengi e Huber (2001), os Cubos de Qualidade estão materializados apenas nos DWs locais, sendo que o nível federado contém apenas os metadados desses cubos.

Quando a ferramenta cliente executa uma consulta federada, ela tem a possibilidade de solicitar também informações sobre a qualidade dos dados. A solicitação é enviada para o Serviço de Qualidade que, conforme descrito na seção 4.3, irá enviar ao Processador

de Consultas uma consulta aos cubos de dados ou aos cubos de qualidade. O papel do Processador de Consultas em relação ao AQUAWARE é análogo ao papel desempenhado pelo Serviço de Consulta na seção 4.3. Tanto para o Processador de Consultas, quanto para os mediadores e os DWs locais não há diferenças se a consulta refere-se aos cubos de dados ou aos Cubos de Qualidade, pois ambos são cubos do DW, armazenados e manipulados de forma semelhante. A diferença entre eles é apenas semântica. Logo, o funcionamento do Processador de Consultas, dos mediadores e dos DWs locais nesta arquitetura é idêntico ao proposto por Mangisengi e Huber (2001). Assim como na seção 4.3, o Serviço de Qualidade funciona como um "cliente" do Processador de Consultas, tornando o acesso aos Cubos de Qualidade transparente para a ferramenta cliente e fornecendo uma interface padrão para o fornecimento de informações sobre a qualidade.

4.5 O METAMODELO DO REPOSITÓRIO DE METADADOS DA QUALIDADE

No AQUAWARE, os objetivos, questionamentos, métricas e medidas da qualidade são armazenados no Repositório de Metadados da Qualidade. O Metamodelo desse Repositório é apresentado nesta seção. A seção 4.5.1 apresenta as necessidades de representação do Metamodelo, que é descrito na seção 4.5.2. A seção 4.5.3 apresenta uma versão simplificada do Metamodelo do RMQ, adaptada ao escopo do presente trabalho

4.5.1 NECESSIDADES DE REPRESENTAÇÃO NO METAMODELO

Para uma melhor compreensão do Metamodelo é preciso conhecer, inicialmente, as necessidades de representação que existem em relação ao RMQ, que são:

- N1: a representação dos diferentes tipos de objetos que fazem parte do ambiente de *Data Warehouse* (fatos, dimensões, tabelas, processos de transformação, etc.);
- N2: a possibilidade de armazenar informações sobre a qualidade de objetos pertencentes a diferentes níveis de abstração da arquitetura de modelagem multicamadas, provendo uma a visão integrada e uniforme desses objetos. Por exemplo,

- no caso da figura 4.7, é necessário armazenar dados sobre a qualidade da variável de fato "Total vendas" e do valor "1000", que é uma instância de Total vendas.
- N3: a necessidade de um mecanismo que possibilite que os objetos do DW sejam referenciados no Metamodelo, para que medidas de qualidade sejam associadas a eles;
- N4: a representação dos relacionamentos entre objetos de diferentes níveis de abstração. Por exemplo, no caso da figura 4.7, o Metamodelo do Repositório deve ser capaz de representar que "1000" é uma instância de "Total_vendas", que por sua vez, é uma instância de "Variável Fato".

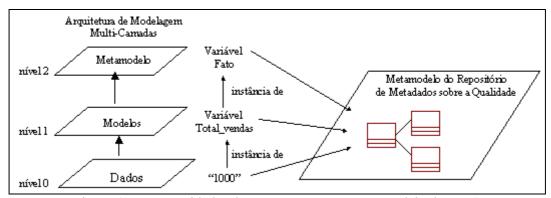


Figura 4.7 – Necessidades de representação no Metamodelo do RMQ

4.5.2 DESCRIÇÃO DO METAMODELO

A seguir é descrito o Metamodelo do RMQ, juntamente com as soluções adotadas para atender as necessidades especificadas na seção 4.5.1. O Metamodelo está ilustrado na figura 4.8.

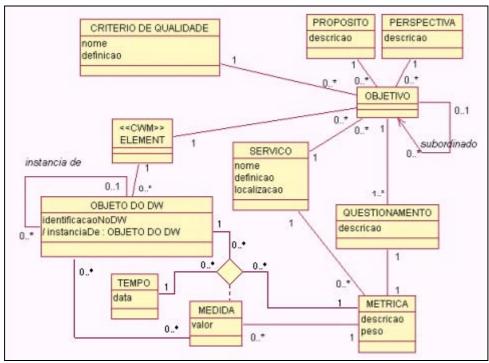


Figura 4.8 – O Metamodelo do Repositório de Metadados da Qualidade

O Metamodelo do Repositório de Metadados da Qualidade foi elaborado como uma extensão do CWM, pela associação de suas classes à classe ELEMENT do CWM. O CWM (OMG, 2001) vai ao encontro da necessidade prevista em N1, pois suas classes representam objetos existentes no ambiente do DW. A classe ELEMENT foi utilizada porque essa é a classe de mais alto nível do CWM e representa, consequentemente, qualquer outra classe deste metamodelo. Apesar de o CWM contemplar a possibilidade de mensuração de características de objetos do modelo, através da classe MEASUREMENT, conforme apresentado no capítulo 2 (seção 2.6), o escopo dessa classe está limitado ao nível 2 da arquitetura de modelagem multi-camadas, não contemplando a mensuração de objetos em todos os níveis. Por este motivo, ela não pôde ser utilizada para medir qualidade no Metamodelo do RMQ.

Os parâmetros de avaliação e as medidas de qualidade foram modelados segundo paradigma GQM (BASILI et al., 1994), com base no Metamodelo da Qualidade no DW de Vassiliadis (VASSILIADIS, 2000), descrito na seção 2.5.8, sobre o qual foram propostas algumas adaptações, com o intuito de atender às necessidades previstas em N2, N3 e N4. Apesar de já prever a mensuração de objetos em diversos níveis, o Metamodelo de Vassiliadis não possibilita uma visão integrada e uniforme desses objetos, além de não representar os relacionamentos entre objetos de diferentes níveis.

As classes PROPOSITO, PERSPECTIVA, CRITERIO DE QUALIDADE, OBJETIVO, QUESTIONAMENTO, METRICA e TEMPO, do Metamodelo do RMQ, têm o mesmo significado que as respectivas classes no Metamodelo de Vassiliadis. As adaptações realizadas no Metamodelo de Vassiliadis são descritas a seguir.

Uma MEDIDA consiste da mensuração de um OBJETO DO DW em relação a uma METRICA num determinado momento (TEMPO). Logo, ela é representada como uma associação ternária entre METRICA, OBJETO DO DW e TEMPO. A classe MEDIDA foi simplificada no Metamodelo do RMQ, para representar apenas os valores reais. Na versão original do Metamodelo de Vassiliadis, essa classe representa também as medidas estimadas. Em alguns casos, o valor da MEDIDA pode variar em função de outros objetos do DW (por exemplo: a qualidade dos fatos pode variar em função das dimensões consideradas). A associação "depende", entre MEDIDA e OBJETO DO DW, representa esse relacionamento.

A classe OBJETO DO DW representa os objetos do ambiente de *Data Warehouse* pertencentes ao níveis 0 e 1. Os objetos do nível 2 são representados pela classe ELEMENT do CWM.

Os objetos do DW são referenciados no Metamodelo através do atributo "identificacaoNoDW", que contém o valor do identificador global único do objeto no *Data Warehouse*, conforme descrito na seção 4.3.

A auto-associação "instância de", em OBJETO DO DW, representa o relacionamento entre os objetos dos níveis 0 e 1. Essa associação tem cardinalidade diferente de zero apenas para os relacionamentos com objetos do nível 0, os quais terão uma referência para os objetos do nível 1. Os objetos do nível 1 referenciam os objetos do nível 2 através da associação entre OBJETO DO DW e ELEMENT, a qual expressa que um OBJETO DO DW é uma instância de ELEMENT.

Por exemplo, no caso da figura 4.7, verifica-se que:

- "1000" é uma instância de "Total_vendas" porque a auto-associação "instância de" em "1000" aponta para "Total_vendas";
- "Total_vendas" é uma instância de ELEMENT (mais especificamente, da subclasse de ELEMENT que representa uma variável de fato, no CWM), pois existe uma

associação entre OBJETO DO DW e ELEMENT. Logo, Total_vendas é uma "variável fato".

• de forma transitiva, obtém-se que "1000" é uma "Variável fato", pois "1000" é uma instância de "Total vendas", que é uma "Variável fato".

Outra característica do Metamodelo do RMQ é a representação dos componentes (pertencentes à Biblioteca de Componentes), que calculam os valores da MEDIDA e do OBJETIVO, o que é feito através da classe SERVICO, que contém o nome, a descrição e a localização do componente.

A seguir é apresentada a definição de cada uma das classes que compõem o Metamodelo do RMQ.

• **OBJETIVO**: Finalidade que se pretende atingir com a avaliação da qualidade. Corresponde ao "Goal" do GQM (BASILI et al., 1994). Um OBJETIVO possui um PROPÓSITO em relação a um CRITERIO DE QUALIDADE, que é aplicado sobre um tipo de OBJETO DO DW, sob uma determinada PERSPECTIVA. Exemplo: avaliar (PROPOSITO) a confiabilidade (CRITERIO DE QUALIDADE) da dimensão Produto (OBJETO DO DW), sob a perspectiva do usuário final (PERSPECTIVA).

O grau no qual o OBJETIVO está sendo atingido, num determinado momento, é obtido através das respostas aos questionamentos. Os questionamentos são respondidos pela mensuração do OBJETO DO DW em relação às METRICAS, que resultam em MEDIDAS. A obtenção o grau no qual o OBJETIVO está sendo atingido, utilizando as MEDIDAS, é realizada por componentes representados pela classe SERVICO.

- **PROPOSITO**: ação que o OBJETIVO propõe realizar. Exemplos: avaliar, aprimorar, prever, controlar, etc.
- **PERSPECTIVA**: ponto de vista sob o qual OBJETIVO foi definido. Exemplos: usuário final, administrador do *Data Warehouse*, etc.
- CRITERIO DE QUALIDADE: critério de qualidade endereçado pelo OBJETIVO.
 Exemplo: confiabilidade, completeza, acurácia, etc.

- QUESTIONAMENTO: esforço que caracteriza a forma pela qual o objetivo será avaliado. Corresponde à "Question" do GQM (BASILI et al., 1994). Procura caracterizar o OBJETO DO DW em relação a um CRITERIO DE QUALIDADE e determinar sua qualidade sob uma determinada PERSPECTIVA. O QUESTIONAMENTO é respondido através de uma METRICA.
- METRICA: caracterização da ação de obtenção de um conjunto de informações sobre um OBJETO DO DW, que respondam ao questionamento quantitativamente.
 A métrica está associada ao SERVIÇO que calcula a medida de um OBJETO DO DW em relação a ela. Corresponde à "Metric" da metodologia GQM.(BASILI et al., 1994).
- **OBJETO DO DW**: corresponde a qualquer objeto do ambiente de *Data Warehouse* (modelado no CWM), sobre o qual são aplicadas medidas de qualidade. O atributo "identificacoNoDW" representa a referência ao objeto real no DW, que é feita através de um identificador global único, conforme descrito na seção 4.3.
- MEDIDA: valor resultante da medição do OBJETO DO DW em relação a uma determinada METRICA, num dado momento (TEMPO). Esse valor é obtido pelo SERVICO que está associado à métrica em questão.
- **SERVICO**: componente que calcula as medidas para cada METRICA, e o grau no qual o OBJETIVO está sendo atingido, para cada OBJETIVO. O SERVICO contém o nome, a descrição e a localização do componente, que faz parte da Biblioteca de Componentes. A Biblioteca de Componentes é detalhada mais adiante, na seção 4.7.
- **TEMPO**: momento em que a MEDIDA foi obtida.
- **ELEMENT**: classe equivalente à classe ELEMENT do CWM.

4.5.3 METAMODELO DO RMQ SOB A PERSPECTIVA DO USUÁRIO FINAL

O Metamodelo do RMQ foi concebido de forma bastante genérica e abrangente, possibilitando o armazenamento de metadados sobre a qualidade de qualquer objeto do DW, pertencente a qualquer nível da arquitetura multi-camadas. Além disso, ele permite a definição dos objetivos da qualidade sob diferentes perspectivas.

Ao abordar, no presente trabalho, os critérios de qualidade sob a perspectiva do usuário final, é possível realizar simplificações nesse Metamodelo para ajustá-lo às reais necessidades do escopo considerado. Sendo assim, é possível afirmar que, no Metamodelo, o PROPOSITO será sempre "avaliar" e a PERSPECTIVA, o "usuário final". Conseqüentemente, o OBJETIVO irá variar apenas em função do CRITERIO DE QUALIDADE e pode ser considerado como tal. A figura 4.9 ilustra o Metamodelo do RMQ sob a perspectiva do usuário final, onde o OBJETIVO é representado como um CRITERIO DE QUALIDADE.

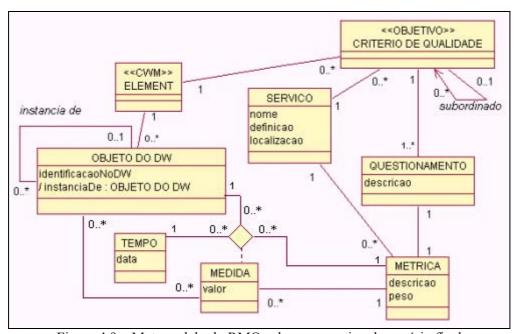


Figura 4.9 – Metamodelo do RMQ sob a perspectiva do usuário final

Outra simplificação inerente ao escopo considerado no presente trabalho diz respeito aos tipos de objeto do DW, cuja qualidade será avaliada. O Metamodelo do RMQ possibilita a definição de objetivos, questionamentos e métricas para qualquer objeto do ambiente de *Data Warehouse*. Contudo, o escopo em questão considera apenas Fatos, Dimensões, Células e Tuplas, uma vez que esses objetos são os mais interessantes sob o ponto de vista do usuário final, conforme identificado no levantamento empírico realizado através de entrevistas, junto a empresas envolvidas em grandes projetos de *Data Warehouse*. Logo, nessa versão do Metamodelo, a classe ELEMENT representa apenas esses quatro tipos de objetos.

4.6 CUBOS DE QUALIDADE

Os Cubos de Qualidade armazenam o grau de qualidade dos objetos do DW e são carregados por componentes presentes na Biblioteca de Componentes, a partir das medidas de qualidade armazenadas no RMQ.

Optou-se pelo armazenamento dos graus de qualidade (objetivos) em Cubos de Qualidade porque, do ponto de vista do levantamento empírico realizado em entrevistas com empresas envolvidas em grandes projetos de DW (tanto do setor público, quanto do privado), ficou evidenciada a necessidade de uma visualização multidimensional da qualidade, que pode variar com o tempo e segundo diferentes perspectivas (ABREU, 2003) (ELIAS, MARTINHO, 2003) (CAMPOS, 2003). Um exemplo são as variáveis de fato, cuja qualidade varia em função da dimensão considerada. A modelagem dessas informações como cubos, agiliza as consultas e favorece a realização de análises segundo as diferentes perspectivas. Assim, foram criados Cubos de Qualidade para objetos pertencentes aos diferentes níveis camadas da arquitetura de modelagem multicamadas. Alguns deles referem-se a objetos pertencentes ao nível 0 (instâncias dos dados), como células e tuplas de tabelas de fatos. Outros referem-se a objetos do nível 1, como dimensões e variáveis de fato.

Os Cubos de Qualidade são construídos segundo a abordagem de modelagem multidimensional, de forma análoga aos cubos de dados, e ficam armazenados no DW junto com esses. A diferença é que ao invés de conterem dados, eles contêm metadados de qualidade. Desta forma, quando o cliente realiza uma consulta, é possível retornar, além dos dados em si, informações sobre a qualidade dos mesmos, através de consultas aos Cubos de Qualidade.

Os Cubos de Qualidade definidos para o AQUAWARE contemplam quatro tipos de Objetos do DW: Dimensões, Fatos, Tuplas e Células. Esses tipos de objeto foram escolhidos por terem sido identificados, no levantamento empírico, como estando entre os mais importantes sob o ponto de vista do usuário final. Desta forma, a modelagem multidimensional dos Cubos de Qualidade contempla a definição de quatro tipos de cubo, descritos a seguir. São eles: o Cubo das Dimensões, o Cubo dos Fatos, Cubos das Tuplas, os Cubos das Células. Existem ainda os Cubos de Anotações, que contêm informações adicionais sobre aqueles.

4.6.1 O CUBO DAS DIMENSÕES

Contém os graus de qualidade das dimensões existentes nos cubos de dados. É composto pela dimensões TEMPO, DIMENSAO e CUBO. A dimensão DIMENSAO representa todas as dimensões dos cubos de dados e a dimensão CUBO, os cubos de dados. A tabela de fatos possui como medidas o graus de qualidade das dimensões. A figura 4.10 apresenta o modelo dimensional do Cubo das Dimensões.

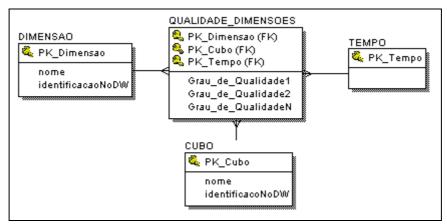


Figura 4.10 – Modelo Dimensional do Cubo das Dimensões

Um exemplo de informação contida nesse cubo é o grau referente à atualidade da dimensão "Loja", no cubo "Vendas", em 11/11/2003.

4.6.2 O CUBO DOS FATOS

Contém os graus de qualidade dos fatos encontrados nos cubos de dados. É composto pelas dimensões TEMPO, FATO, CUBO e GRUPOS_DIMENSAO. A dimensão FATO representa todos os fatos dos cubos dos dados e a dimensão CUBO, os cubos de dados.

Ao contrário das dimensões, a qualidade dos fatos não pode ser analisada de forma isolada, pois o grau de qualidade de uma variável de fato pode variar em função das dimensões consideradas. Para representar as dimensões segundo as quais a qualidade do fato foi analisada, o Cubo dos Fatos possui um relacionamento muitos-para-muitos com a dimensão DIMENSAO, descrita na seção anterior. Este é um caso de dimensão muitos-para-muitos (KIMBALL et al., 1998), que foi resolvido com a utilização de uma "tabela ponte" (*bridge table*) (KIMBALL et al., 1998) entre a tabela de fatos a dimensão DIMENSAO, denominada GRUPO DIMENSAO.

A tabela de fatos possui como medidas o graus de qualidade para os fatos. A figura 4.11 apresenta o modelo dimensional do Cubo dos Fatos.

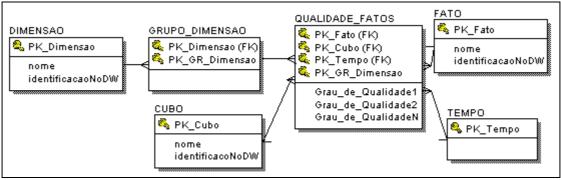


Figura 4.11 – Modelo Dimensional do Cubo dos Fatos

Um exemplo de informação contida nesse cubo é o grau referente à atualidade da variável de fatos "Total_Vendas", no cubo "Vendas", sob a perspectiva da dimensão "Loja", em 11/11/2003.

4.6.3 OS CUBOS DE TUPLAS

Os Cubos das Tuplas foram definidos como espécies de "tabelas de auditoria" para cada tabela de fatos, com o objetivo de armazenar o grau de qualidade de suas tuplas. A tabela de auditoria está associada às mesmas dimensões que a tabela de fatos, possibilitando que a ligação entre os dados e seus respectivos metadados de qualidade seja feita através de uma operação do tipo *drill-across*⁴. Se a dimensão TEMPO não estiver relacionada à tabela de fatos, pode-se criar uma associação da mesma com a tabela de auditoria. O Cubo de Tuplas possui como medidas, os graus de qualidade das tuplas da tabela de fatos a que se refere. A figura 4.12 apresenta o modelo dimensional das tabelas de auditoria que compõem os Cubos de Tuplas.

⁴ Processo de efetuar uma junção entre duas tabelas de mesma granularidade e que estejam ligadas ao mesmo conjunto de dimensões (KIMBALL, 1998).

_

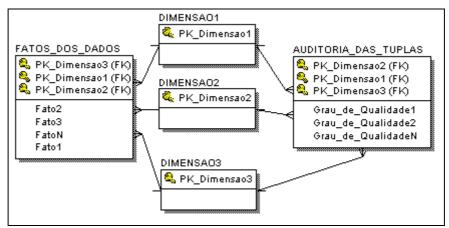


Figura 4.12 - Modelo Dimensional de um Cubo de Tuplas

Um exemplo de informação contida nesse cubo é o grau referente à atualidade de uma tupla (registro) da tabela de fatos "Vendas", em 11/11/2003.

4.6.4 OS CUBOS DE CÉLULAS

Os Cubos das Células foram definidos como espécies de "tabelas de auditoria" para cada tabela de fatos, com o objetivo de armazenar o grau de qualidade de suas células. A célula corresponde ao valor de uma variável de fato, numa determinada tupla.

A tabela de auditoria está associada às mesmas dimensões que a tabela de fatos e também à dimensão FATO. A associação com a dimensão FATO tem como objetivo identificar a que variável de fato da tupla (célula) os graus de qualidade se referem. Assim como nos Cubos das Tuplas, a ligação entre os dados e seus respectivos metadados pode ser feita através de uma operação do tipo *drill-across*. Se a dimensão TEMPO não estiver relacionada à tabela de fatos, pode-se criar uma associação da mesma com a tabela de auditoria. A tabela de auditoria possui como medidas os graus de qualidade das células da tabela de fatos a que se refere. A figura 4.13 apresenta o modelo dimensional das tabelas de auditoria que compõem os Cubos de Células.

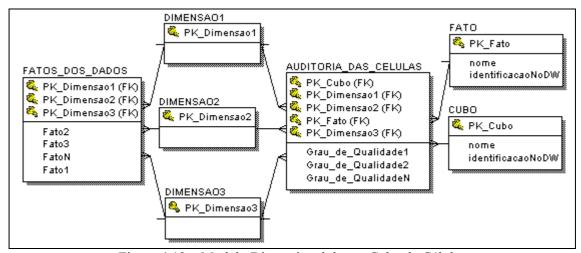


Figura 4.13 – Modelo Dimensional de um Cubo de Células

Um exemplo de informação contida nesse cubo é o grau referente à atualidade da célula que contém o valor "100" para o fato "Total_Vendas", no cubo "Vendas", em 11/11/2003.

4.6.5 OS CUBOS DE ANOTAÇÕES

Os Cubos de Anotações são Cubos de Qualidade que contêm informações não estruturadas e semi-estruturadas, sobre a qualidade. É o caso de informações essencialmente subjetivas, que não são possíveis de serem quantificadas (e por este motivo devem ser apresentadas de forma descritiva), mas que são importantes sob o ponto de vista da qualidade.

Para cada um dos demais Cubos de Qualidade, existe um Cubo de Anotações associado às mesmas dimensões que sua tabela de fatos. A tabela de fatos do Cubo de Anotação contém apenas um campo de texto longo, onde é armazenada a anotação. A ligação entre um Cubo de Qualidade e seu respectivo Cubo de Anotações é feita através de operações do tipo *drill-across*.

Foram definidos quatro Cubos de Anotações, a saber: Cubo de Anotação das Dimensões, Cubo de Anotação dos Fatos, Cubo de Anotação das Tuplas e Cubo de Anotação das Células. As figuras 4.14 a 4.17 ilustram a modelagem dimensoional desses cubos.

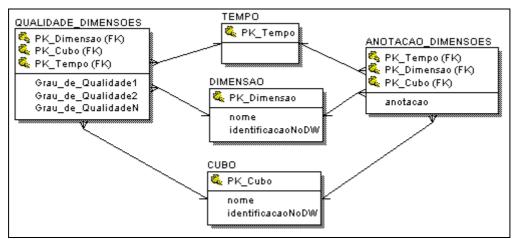


Figura 4.14 – Modelo Dimensional do Cubo de Anotações das Dimensões

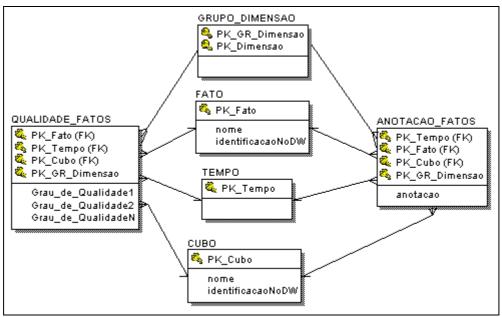


Figura 4.15 – Modelo Dimensional do Cubo de Anotações dos Fatos

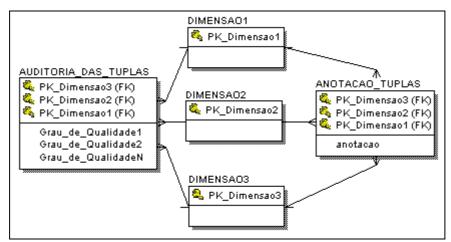


Figura 4.16 – Modelo Dimensional do Cubo de Anotações de Tuplas

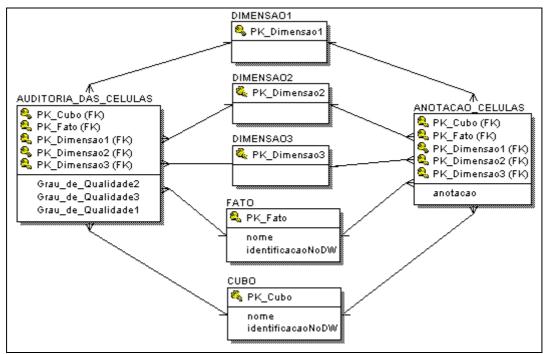


Figura 4.17 – Modelo Dimensional do Cubo de Anotações de Células

Conforme citado anteriormente, a anotação é composta por informações textuais não estruturadas ou semi-estruturadas. Neste último caso, elas podem estar organizadas num documento XML, onde cada elemento representa o grau de qualidade de um critério, contendo as anotações referentes a ele. Abaixo é apresentado um exemplo de XML *Schema* que pode ser utilizado como modelo para a estrutura do campo anotação.

```
<?xml version="1.0" encoding="utf-16"?>
<xs:schema xmlns=""</pre>
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="GrausDeQualidade">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="GrauDeQualidade1" type="xs:string"</pre>
         minOccurs="0" />
        <xs:element name="GrauDeQualidade2" type="xs:string"</pre>
         minOccurs="0" />
        <xs:element name="GrauDeQualidadeN" type="xs:string"</pre>
         minOccurs="0" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

4.7 BIBLIOTECA DE COMPONENTES

A Biblioteca de Componentes é composta pelos artefatos de *software* que carregam o modelo de qualidade no RMQ e implementam os processos de medição da qualidade, cálculo dos objetivos, carga dos Cubos de Qualidade e fornecimento de informações sobre a qualidade dos dados para os clientes. Esses componentes podem ser classificados da seguinte forma:

- Componentes que carregam o modelo que qualidade no RMQ;
- Componentes que geram as medidas de qualidade;
- Componentes que calculam dos objetivos;
- Componentes que fazem a carga dos Cubos de Qualidade; e
- Componentes do Serviço de Qualidade

Esses tipos de componentes são descritos a seguir. Para que cada componente possa ser identificado de forma única, será utilizada a seguinte convenção para a nomeação dos mesmos: o nome inicia com um código, que começa pela letra C (inicial de componente), seguida de um número seqüencial. O anexo III apresenta a relação de todos os componentes da Biblioteca e o anexo IV, a especificação dos mesmos.

4.7.1 CARGA DO MODELO DE QUALIDADE NO RMQ

Consiste na inclusão dos objetivos, questionamentos, métricas e na carga dos objetos do DW no RMQ. Os objetivos, questionamentos e métricas são incluídos pelo administrador do DW, utilizando o componente C01 – Inclusão dos Objetivos, Questionamentos e Métricas. Os objetos do DW são incluídos de forma automática, pelo componente C02 – Carga dos Objetos do DW no RMQ. O anexo IV contém as especificações desses componentes.

4.7.2 GERAÇÃO DAS MEDIDAS DE QUALIDADE

Consiste na obtenção das medidas de qualidade para os objetos do DW, a partir das métricas definidas para os mesmos, seguida do seu armazenamento no Repositório de Metadados da Qualidade.

Uma característica do processo de medição da qualidade é que ele não é único. A forma de medir a qualidade de um objeto em relação a uma determinada métrica varia em função do tipo do objeto que está sendo medido (*Element* do CWM). Logo, para cada par "Métrica – Tipo de Objeto", existirá um componente específico, que implementa o processo de medição da qualidade, e apresenta o valor gerado como retorno.

Outra característica dos processos de medição da qualidade é que eles são, em geral, executados de forma automática pelos componentes, que extraem, do próprio ambiente do DW, as informações necessárias para a geração dos resultados. Contudo, nem sempre é possível executar a medição de forma totalmente automatizada. Conforme citado no capítulo 2, há situações em que a avaliação da qualidade envolve questões de caráter extremamente subjetivo, sendo necessária a intervenção humana para a execução de partes do processo.

Os componentes responsáveis pela geração das medidas de qualidade devem ser implementados à medida em que forem definidas as métricas para cada tipo de objeto do DW, no RMQ. A classe Serviço contém a descrição e localização de cada componente, relacionando-o ao par "Métrica – Tipo de Objeto" a que ele se refere.

No capítulo 5 são definidas métricas para a mensuração da qualidade de fatos e dimensões, segundo o paradigma GQM (BASILI et al., 1994). O anexo IV contém as especificações dos componentes que realizam a geração das medidas de qualidade em relação a essas métricas.

4.7.3 CÁLCULO DOS OBJETIVOS

Cálculo do grau no qual os objetivos da qualidade estão sendo atingidos, para cada objeto do DW. O cálculo do objetivo é feito a partir das medidas de qualidade do objeto do DW ao qual se refere, as quais ficam armazenadas no RMQ. Contudo, como essas medidas estão em unidades distintas é necessário uniformizá-las para poder somá-las entre si, obtendo assim um valor global para o objetivo. Esse tipo de problema pode ser resolvido pelo uso de métodos de tomada de decisão com múltiplos atributos (NAUMANN,1998) (ZHU, 2002). Para o cálculo dos objetivos é utilizado o método de Pesagem Aditiva Simples (SAW - *Simple Additive Weighting*) (NAUMANN, 1998), descrito a seguir.

As medidas de qualidade são individualmente uniformizadas através das equações (1) e (2), onde a primeira será utilizada para medidas positivas (quanto maior for o valor da medida, maior será a qualidade) e a segunda para medidas negativas (quanto maior for o valor da medida, menor será a qualidade).

$$v_{ij} = \frac{c_{ij} - c_{j}^{min}}{c_{j}^{max} - c_{j}^{min}} \quad (1) \qquad v_{ij} = \frac{c_{j}^{max} - c_{ij}}{c_{j}^{max} - c_{j}^{min}} \quad (2)$$

onde: v_{ij} é o valor uniformizado da medida j para o objetivo i c_{ij} é o valor original da medida j para o objetivo i c_j^{max} é o maior valor que a medida j pode assumir c_j^{min} é o menor valor que a medida j pode assumir

Uma vez aplicadas as funções de uniformização, cada medida que qualidade uniformizada vij terá um valor no intervalo $0 \le v_{ij} \le 1$. Essa propriedade assegura a possibilidade de comparação entre medidas distintas. Assim, poderão ser somadas todas as medidas de qualidade de um mesmo objetivo.

A cada medida de qualidade uniformizada é atribuído um valor de peso que reflete a importância de uma métrica em relação a outras. O vetor de pesos W = (w1, w2, ..., wn) obedece a propriedade $\sum_{j=1}^{n} wj = 1$. O cálculo do objetivo é calculado pela soma ponderada definida na equação (3).

Objetivo_i =
$$\sum_{j=1}^{n} w_j \times v_{ij}$$
 (3)

O Cálculo do Objetivo é implementado pelo componente C03 – Cálculo do Objetivo, cuja especificação encontra-se no anexo IV.

4.7.4 CARGA DOS CUBOS DE QUALIDADE

Consiste no armazenamento dos graus de qualidade dos objetos do DW nos Cubos de Qualidade, exceto para os Cubos de Anotações. Devido à natureza essencialmente subjetiva das informações de qualidade contidas nesses últimos, não é possível realizar a geração automática de seus dados, cuja disponibilização depende da intervenção humana. Assim, as anotações são carregadas de forma independente dos componentes, a partir de arquivos texto XML ou outro formato compatível.

No processo de carga , o componente C03 (Cálculo do Objetivo) é utilizado para a geração dos graus de qualidade, a partir das medidas de qualidade armazenadas no RMQ. A tabela 4.2 apresenta os nomes componentes responsáveis pelos processos de carga e o anexo IV, a sua especificação.

Componente	Nome	
C04	Carga da dimensão DIMENSAO	
C05	Carga da dimensão TEMPO	
C06	Carga da tabela de fatos do Cubo das Dimensões	
C07	Carga da dimensão GRUPO_DIMENSAO	
C08	Carga da dimensão FATO	
C09	Carga da tabelas de fatos do Cubo dos Fatos	
C10	Carga da dimensão CUBO	
C11	Carga da tabela de auditoria das tuplas	
C12	Carga da tabela de auditoria das células	

Tabela 4.2 – Componentes para a carga dos Cubos de Qualidade

4.7.5 OS SERVIÇOS DE CONSULTA E QUALIDADE

Para atender a algumas necessidades inerentes ao contexto interorganizacional, como heterogeneidade, complexidade e necessidade de padronização optou-se pela disponibilização dos componentes que implementam os Serviços de Consulta e Qualidade, como serviços *Web*. Em ambientes analíticos interorganizacionais, uma boa opção são as tecnologias que permitem a realização de análises em ambientes heterogêneos, utilizando a *Web* como infra-estrutura de comunicação. Neste caso o padrão XML Analysis e a API Java OLAP Interface atendem às necessidades aqui apresentadas. Optou-se pela utilização do XMLA por ele não estar atrelado a uma linguagem de programação ou modelo de objetos específicos, além de possuir uma linguagem padrão para consultas multidimensionais, o que é não é trivial de se fazer no JOLAP, pois é realizado através de linhas de comandos que utilizam vários métodos e classes.

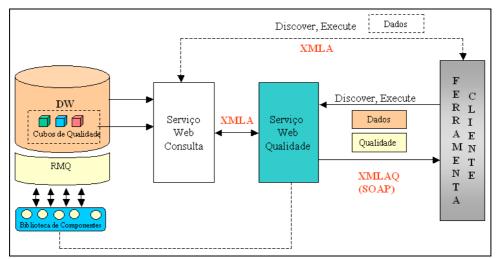


Figura 4.18 – Arquitetura do AQUAWARE utilizando XMLA

Na arquitetura do AQUAWARE utilizando XMLA, ilustrada na figura 4.18, tanto o Serviço de Consulta, quanto o Serviço de Qualidade são disponibilizados como serviços Web. O AQUAWARE não contempla a especificação do Serviço de Consulta. Este, corresponde a qualquer serviço que implemente a funcionalidade de realização de consultas em ambientes analíticos, utilizando o padrão XMLA para a troca de mensagens. Nele, o método Discover do XMLA é utilizado para recuperar informações sobre a estrutura do DW (metadados) e o método Execute, para executar consultas. O anexo I contém a descrição dos métodos Discover e Execute do XMLA (MICROSOFT, HYPERION, 2002).

O Serviço de Qualidade utiliza uma versão adaptada do XMLA, aqui denominada XMLA-Q, onde os métodos *Discover* e *Execute* foram alterados para contemplar funcionalidades relativas à qualidade. O XMLA-Q utiliza mensagens no formato SOAP para a comunicação com os clientes. O componente do Serviço de Qualidade implementa os métodos do XMLA-Q, fazendo o tratamento das diferenças entre o XMLA-Q e o XMLA. Assim, a comunicação com o Serviço de Consulta pode ser feita utilizando-se exclusivamente o padrão XMLA original. No anexo V são apresentados os métodos *Discover* e *Execute* do XMLA-Q, implementados pelo componente que disponibiliza o Serviço de Qualidade (C13).

CAPÍTULO 5

INDICADORES DA QUALIDADE DE DADOS NO DW SOB A PERSPETIVA DO USUÁRIO FINAL

Conforme apresentado no capítulo anterior, o conceito de qualidade foi modelado no Repositório de Metadados da Qualidade do AQUAWARE sob a forma de objetivos, questionamentos e métricas, segundo o paradima GQM (BASILI, 1994). Esse repositório é flexível no sentido que permite a definição de diferentes indicadores de qualidade (expressos como objetivos, questionamentos e métricas), particulares de cada situação específica, através de um processo de instanciação do seu metamodelo e da associação das métricas aos componentes que medem a qualidade dos objetos do DW segundo as mesmas.

Neste capítulo são definidos indicadores para avaliar a qualidade de dados sob a perspectiva do usuário final, considerando as particularidades do contexto interorganizacional. Inicialmente, na seção 5.1, foram selecionados alguns critérios de qualidade que serviram como ponto de partida para a identificação dos indicadores. Na seção 5.2, o paradigma GQM (BASILI et al., 1994) foi utilizado para orientar a definição dos mesmos. Os tipos de objetos do DW avaliados pelos indicadores são as variáveis de fato (medidas) e as dimensões. As especificações dos componentes responsáveis pela geração das medidas de qualidade para as métricas definidas na seção 5.2 são apresentadas no anexo IV. Finalmente, a seção 5.3 traz algumas considerações sobre o processo de definição dos indicadores de qualidade.

5.1 SELEÇÃO DOS CRITÉRIOS DA QUALIDADE

Apesar de existirem diversas propostas de critérios para a avaliação da qualidade de dados, conforme visto no capítulo 2, ainda não há um consenso em relação ao conjunto de critérios que melhor expressa o conceito de qualidade no âmbito dos dados. As divergências também ocorrem em relação ao significado, pois um mesmo critério pode ser encontrado na literatura com diferentes definições. Por este motivo, ao referenciar um critério, é importante explicitar também o seu significado.

Para a avaliação da qualidade de dados no AQUAWARE foi selecionado um conjunto de critérios, a partir dos quais foram definidas métricas para a mensuração da qualidade utilizando a abordagem GQM (BASILI et al., 1994), descrito na seção 2.1.1. Esta seção apresenta e descreve os critérios selecionados. As métricas obtidas são apresentadas na seção 5.2 e as especificações dos componentes responsáveis pela geração das medidas de qualidade para as mesmas são apresentadas no anexo IV.

A identificação dos critérios foi realizada considerando-se os seguintes aspectos, inerentes ao escopo do presente trabalho:

- perspectiva do usuário final;
- particularidades do ambiente de Data Warehouse; e
- contexto interorganizacional.

Essa escolha teve como base:

- os critérios orientados aos consumidores de dados, propostos por Wang et al.(1996)
 e Wand e Wang (1996);
- os critérios para a qualidade em contextos interorganizacionais, definidos por Bertolazzi e Scannapieco (2001) e Aubert et al. (2003);
- os requisitos para a qualidade na integração de dados, definidos por Aubert et al.
 (2003); e
- os critérios identificados em entrevistas realizadas em empresas com diferentes perfis: uma empresa atacadista da indústria de bebidas, uma empresa do setor público e um fabricante de ferramenta OLAP e fornecedor de soluções na área de inteligência de negócio (ABREU, 2003) (CAMPOS, 2003) (ELIAS, MARTINHO, 2003).

Os critérios selecionados são apresentados a seguir.

Acurácia

Representa a distância entre dois valores v e v', sendo v considerado o valor correto (WANG et al., 1996)(WAND, WANG, 1996). Ela pode ser classificada em três tipos: acurácia sintática, acurácia semântica e acurácia do conteúdo.

A acurácia sintática diz respeito às regras de formação sintática, ao tipo e ao domínio dos dados. As diferenças ocorrem porque os sistemas que atuam como origem dos dados podem ter sido desenvolvidos em organizações diferentes, com base em diferentes requisitos e em momentos diferentes, para dar suporte a processos de negócio específicos de cada uma delas. Conseqüentemente, eles podem apresentar diferenças quanto às regras sintáticas. Por exemplo, o tipo de um determinado elemento de dado numa organização A pode ser diferente do tipo para o elemento de dado equivalente numa organização B (em A, um código pode conter apenas inteiros, enquanto que em B, esse mesmo código pode conter inteiros e caracteres alfabéticos).

A acurácia semântica diz respeito ao significado do dado, à integridade referencial e às regras de negócio. Assim como conceitos podem ter significados diferentes em função do contexto considerado, os mesmos dados podem ter significados diferentes, dependendo da organização à qual pertençam. As regras semânticas para um determinado elemento de dados, numa organização A, podem ser diferentes das regras semânticas para o elemento de dado equivalente, numa organização B. Por exemplo, a fórmula do cálculo da receita na organização A pode ser diferente na organização B. Se os dados de receita de A e B forem agregados no DW, o resultado produzido não fará sentido.

Em relação à integridade referencial, a ocorrência, na tabela de fatos, de um código de dimensão que não esteja cadastrado, pode acarretar problemas durante a consolidação das informações, pois esse código não possuirá um nível na hierarquia da dimensão e, consequentemente, os fatos associados a ele aparecerão como não classificados.

A acurácia do conteúdo diz respeito ao grau no qual o conteúdo do dado está correto em relação à realidade. Geralmente, esse critério é avaliado através da comparação do dado com uma fonte conhecida, cuja correção é garantida. Apesar de sua relevância para o usuário final, a avaliação desse critério não é simples de ser obtida, pois na maioria dos casos não existe uma fonte conhecida para ser utilizada na comparação. Um

exemplo onde isso é possível é no caso de "Preços de Produtos", onde as informações de preços no DW podem ser comparadas com tabelas de preços externas. Outros exemplos são as informações sobre níveis de estoque que, via de regra, não podem assumir valores negativos.

Precisão

Grau no qual o valor do dado corresponde a um valor aproximado em relação ao valor real (BERTOLAZZI, SCANNAPIECO, 2001)(WANG et al., 1996)(WAND, WANG, 1996). Este critério difere da acurácia de conteúdo porque, neste caso, a diferença entre o dado e o valor real não ocorre devido a erros e sim a aproximações realizadas de forma proposital. É o caso, por exemplo dos valores estimados (WAND, WANG, 1996). São exemplos também os casos em que se retiram os centavos de valores monetários. Um outro exemplo, relativo a informações de preços de produtos, ocorre com empresas que atuam como fornecedoras de produtos, cujo preço de venda varia em função do distribuidor. A variação ocorre porque alguns distribuidores têm condição de praticar preços mais baixos que outros. Para realizar a consolidação das vendas no DW é necessário fixar um preço para o produto, o que pode ser feito através de cálculos para a obtenção de um preço médio. Nesse caso, a informação de preço não reflete um valor real específico e sim um valor médio, obtido a partir dos valores reais.

Completeza

É o grau no qual os elementos de um esquema estão presentes nas instâncias (AUBERT et al, 2003)(BERTOLAZZI, SCANNAPIECO, 2001)(WANG et al., 1996)(WAND, WANG, 1996). No DW, os problemas de completeza podem ser causados por diversos motivos como:

- Carga incompleta na tabela de fatos: por exemplo, numa tabela de fatos de "Vendas"
 pode-se ter feita a carga para todas as lojas, exceto para a filial de "Niterói".
- Dimensões incompletas: por exemplo, se uma dimensão contendo "Promoções Semanais" não for atualizada semanalmente, poderá haver tuplas na tabela de fatos de "Vendas" cujo código de promoção não está cadastrado nessa dimensão. Neste

caso, a dimensão incompleta estará acarretando problemas na integridade referencial da tabela de fatos em relação à mesma.

Outro aspecto importante em relação à completeza é a forma como são tratados os valores nulos. O não preenchimento de um dado pode ocorrer por diversas razões: por não estar disponível, por não ser aplicável, por não existir valor no domínio que o classifique corretamente ou por ser desconhecido. A existência de códigos especiais para indicar o motivo da ausência de valor para o dado é um fator importante que deve ser considerado na avaliação da completeza.

Em alguns casos, é interessante disponibilizar, além do grau de completeza, uma anotação explicitando o tipo de informação que está ausente. Por exemplo, se o valor agregado das vendas para um determinado dia não inclui uma determinada loja, é importante apresentar para o usuário, uma anotação que transmita essa informação.

Atualidade

É a diferença de tempo entre o instante em que o dado é atualizado na origem e o instante em que ele é disponibilizado no DW (BERTOLAZZI, SCANNAPIECO, 2001)(WANG et al., 1996)(WAND, WANG, 1996). É o caso de informações que podem variar com o tempo, como dados de endereço, por exemplo.

Se a carga das informações dos sistemas transacionais no DW não for realizada periodicamente, segundo intervalos de tempo pré-definidos, as consultas realizadas no DW podem apresentar defasagens significativas em relação à realidade. Uma boa prática é a definição de um calendário de atualizações, especificando a periodicidade das cargas, e a utilização de um *log* das atualizações efetivadas no DW.

Ainda que as cargas sejam executadas dentro da periodicidade planejada, existe o risco de alguma informação não ser carregada por estar indisponível. Por exemplo, se a informação de custo unitário do produto para o mês corrente não estiver disponível e for utilizado o valor de custo do mês anterior, uma possível tabela de fatos sobre "Custo dos Produtos" estará comprometida quanto à atualidade.

Reputação

Corresponde à credibilidade da organização de origem, que atua como fonte dos dados do DW (BERTOLAZZI, SCANNAPIECO, 2001)(WANG et al., 1996). Esse critério possui especial relevância no contexto interorganizacional, onde a origem da informação pode ser externa à organização. Contudo, medir a credibilidade de uma fonte de dados não é uma tarefa simples. A obtenção de um valor para o grau de credibilidade envolve uma série de fatores de caráter subjetivo, que dependem do ponto de vista de quem está analisando, sendo que dificilmente seria possível chegar-se a resultados factíveis sem a intervenção humana. Ainda assim os resultados seriam extremamente dependentes da opinião particular do avaliador. Desta forma, considerase que, ao invés de se tentar obter graus para esse critério, é mais adequado apresentar para o usuário a descrição das fontes dos dados. Esse tipo de informação pode ser classificado como não estruturada ou semi-estruturada, e deve ser disponibilizada no AQUAWARE como uma anotação (ROTHEMBERG, 1996).

Alterações Propositais

Grau de alteração proposital nos dados. É o caso, por exemplo, de informações interorganizacionais que, por motivos políticos, não podem ser apresentadas com seu valor real. Aqui, assim como no caso da reputação dos dados, a utilização de uma anotação fornecendo explicações sobre as alterações realizadas é considerada mais adequada do que a obtenção de graus.

Confiabilidade

A confiabilidade expressa o grau de confiança que usuários têm nos dados (WANG et al., 1996). O conceito de confiabilidade é complexo e pode envolver outros critérios como acurácia, reputação, completeza, etc. Contudo, uma definição mais objetiva desse critério apóia-se no fato de que a confiança do usuário é diretamente proporcional à capacidade do DW de refletir corretamente os dados dos sistemas transacionais. Se o usuário realiza uma consulta no DW, e em seguida consulta as mesmas informações nos sistemas transacionais e obtém resultados diferentes, ele começa a ficar desconfiado. Se este tipo de situação ocorre repetidas vezes, ele perde a confiança no DW e pára de utilizá-lo. Assim, a confiabilidade pode ser avaliada através de comparações entre os

dados carregados no DW e os dados existentes nos sistemas de transacionais. Contudo, essa comparação só poderá ser realizada se os sistemas de origem pertencerem à organização ou, caso contrário, se a mesma tiver acesso a eles.

5.2 DEFINIÇÃO DOS INDICADORES (MÉTRICAS)

Nesta seção são definidas as métricas, aqui denominadas indicadores de qualidade, para os critérios apresentados na seção 5.1. A metodologia GQM (BASILI et al., 1994), descrita na seção 2.1.1, foi utilizada para a obtenção das mesmas. Destaca-se que o levantamento realizado não teve como objetivo identificar métricas para todos os tipos de objetos do DW. A idéia aqui é apresentar um conjunto de métricas para alguns tipos de objeto, a saber os fatos e as dimensões, de forma que possam ser definidos componentes para a medição da qualidade dos mesmos. Conforme descrito no capítulo 4, esses componentes irão gerar as medidas de qualidade que serão armazenadas no RMQ e serão utilizadas como base para o cálculo dos graus de qualidade (que correspondem aos objetivos) dos fatos e dimensões. Os graus de qualidade, por sua vez, serão armazenados nos cubos de qualidade.

Conforme visto na seção 4.5.3 de acordo com o escopo considerado no presente trabalho, o objetivo terá sempre como perspectiva, o usuário final, e como propósito, avaliar. Assim, tanto para os fatos quanto para as dimensões, o objetivo irá variar apenas em função do critério de qualidade avaliado. A seguir, são apresentados os objetivos, questionamentos e métricas para objetos do tipo dimensão e fato. Para que esses objetivos, questionamentos e métricas possam ser identificados de forma única, será utilizada a seguinte convenção para a nomeação dos mesmos: o nome do objetivo inicia com um código, que começa pela letra O (inicial de objetivo), seguida de um número seqüencial; o nome do questionamento inicia com um código, que começa pela letra Q (inicial de métrica), seguida do número do objetivo e de um número seqüencial; o nome da métrica inicia com um código, que começa pela letra M (inicial de métrica), seguida do número do questionamento e de um número seqüencial. A métrica retornará sempre um número, pertencente ao intervalo de zero a cem. Na maioria dos casos, optou-se pela utilização de medidas positivas, ou seja, quanto maior for o valor da medida, maior será a qualidade.

5.2.1 INDICADORES PARA OS FATOS

Nesta seção são apresentados os objetivos, os questionamentos e as métricas para objetos do tipo "variável de fato", tendo em vista os seguintes critérios de qualidade: acurácia sintática, semântica e de conteúdo, precisão, completeza, atualidade, reputação, alterações propositais e confiabilidade. A tabela 5.1 apresenta esses indicadores. Os símbolos (+) e (-) ao lado do nome da métrica indicam se a mesma é positiva ou negativa, respectivamente.

Objetivo	Questionamento	Métrica
01	Q1.1	M1.1.1 (+)
Acurácia Sintática	Obedece às regras de formação sintática?	Σ (si * wi) / número total de regras sintáticas
		Onde:
		i= regra de formação sintática
		si = % do número de registros que obedecem à regra de formação sintática i
		wi = peso da regra i
		O vetor de pesos $W = (w1, w2,, wn)$ obedece a propriedade $\Sigma nj=1$ $wj=1$
	Q1.2	M1.2.1 (+)
	Pertence ao domínio de dados?	% de registros que pertencem ao domínio
	Q1.3	M1.3.1 (+)
	Pertence ao tipo de dados?	% de registros que pertencem ao tipo de dados
O2	Q2.1	M2.1.1 (+)
Acurácia Semântica	Obedece às regras de negócio?	Σ (ei * wi) / número total de regras de negócio
		i= regra de negócio
		ei = % de registros que obedecem à regra de negócio i
		wi = peso da regra i
		O vetor de pesos W = (w1, w2,, wn) obedece a propriedade Σ nj=1 wj = 1
	Q2.2	M2.2.1 (+)
	As chaves estrangeiras existem nas dimensões ?	% de registros contendo chaves estrangeiras cadastradas nas dimensões
03	Q3.1	M3.1.1 (+)
Acurácia de Conteúdo	Quando comparado com fonte conhecida, apresenta	% de registros que estão em conformidade

Objetivo	Questionamento	Métrica
	valor correto?	com fonte conhecida
O4 Precisão	Q4.1 Quão próximo do valor real está o valor armazenado?	M4.1.1 (+)
		% de proximidade do valor armazenado em relação ao valor real
O5	Q5.1	M5.1.1 (+)
Completeza	O campo, quando aplicável, está preenchido?	% dos registros onde o campo está preenchido ou está identificado como não aplicável
	Q5.2 Os membros existentes nas dimensões estão cadastrados nas mesmas?	M5.2.1 (+) % dos membros existentes que estão cadastrados nas dimensões
	Q5.3	M5.3.1 (+)
	Os registros foram carregados na tabela de fatos?	% do total de registros que foram carregados com sucesso na tabela de fatos
	Tatios:	Obs: Aqui pode-se utilizar uma anotação que descreva de forma sucinta os registros ausentes.
06	Q6.1	M6.1.1 (+)
Atualidade	Os registros estão sendo carregados com a freqüência esperada?	Valor zero ou cem, correspondendo a desatualizado ou atualizado, respectivamente.
		Obs: Aqui pode-se utilizar uma anotação contendo a data da última atualização.
O7 Reputação	Q7.1	M7.1.1 (+)
Reputação	Qual a origem dos dados?	Anotação com a descrição das fontes dos dados.
08	Q8.1	M8.1.1
Alterações Propositais	As informações foram alteradas propositalmente?	Anotação fornecendo explicações sobre as alterações realizadas.
O9	Q9.1	M9.1.1 (+)
Confiabilidade	Reflete corretamente os dados dos sistemas transacionais?	Utiliza rotinas de verificação, que comparam os dados no DW (destacando-se os agregados) com os dados nos sistemas transacionais.
		A métrica corresponde ao percentual das rotinas que retornaram sucesso (ou seja, o DW reflete corretamente os dados do sistema transacional).
		Obs: Essa verificação só é possível quando o sistema transacional está disponível para consulta. Em contextos interorganizacionais ele pode pertencer a uma organização externa e não estar disponível. Nesse caso, não será possível utilizar essa métrica.

Tabela 5.1 – Indicadores para os Fatos

5.2.2 INDICADORES PARA AS DIMENSÕES

Nesta seção são apresentados os objetivos, os questionamentos e as métricas para objetos do tipo "dimensão", tendo em vista os seguintes critérios de qualidade: acurácia sintática e semântica, completeza, atualidade, reputação e alterações propositais. A tabela 5.2 apresenta esses indicadores. Os símbolos (+) e (-) ao lado do nome da métrica indicam se a mesma é positiva ou negativa, respectivamente.

Objetivo	Questionamento	Métrica
O10	Q10.1	M10.1.1 (+)
Acurácia Sintática	Obedece às regras de formação sintática?	(Σ si * wi) / número total de regras sintáticas
		Onde:
		i= regra de formação sintática
		si = % do número de registros que obedecem à regra de formação sintática i
		wi = peso da regra i
		O vetor de pesos $W = (w1, w2,, wn)$ obedece a propriedade $\Sigma nj=1$ $wj=1$
	Q10.2 Pertence ao domínio de dados?	M10.2.1 (+)
		% de registros que pertencem ao domínio
	Q10.3	M10.3.1 (+)
	Pertence ao tipo de dados?	% de registros que pertencem ao tipo de dados
O11	Q11.1 Obedece às regras de negócio?	M11.1.1 (+)
Acurácia Semântica		(Σ ei * wi) / número total de regras de negócio
		i= regra de negócio
		ei = % de registros que obedecem à regra de negócio i
		wi = peso da regra i
		O vetor de pesos $W = (w1, w2,, wn)$ obedece a propriedade $\Sigma nj=1$ $wj=1$
O12	Q12.1	M12.1.1 (+)
Completeza	Os membros existentes nas dimensões estão cadastrados nas mesmas?	% dos membros existentes que estão cadastrados nas dimensões
	Q12.2 Os membros estão categorizados nas hierarquias?	M12.2.1 (-)
		% dos membros não categorizados nas hierarquias

Objetivo	Questionamento	Métrica
O13	Q13.1	M13.1.1 (+)
Atualidade	Os registros estão sendo carregados com a frequência esperada?	Valor zero ou cem, correspondendo a desatualizado ou atualizado, respectivamente.
		Obs: Aqui pode-se utilizar uma anotação contendo a data da última atualização.
O14	Q14.1	M14.1.1 (+)
Reputação	Qual a origem dos dados?	Anotação com a descrição das fontes dos dados.
O15	Q15.1	M15.1.1 (+)
Alterações Propositais	As informações foram alteradas propositalmente?	Anotação fornecendo explicações sobre as alterações realizadas.

Tabela 5.2 – Indicadores para as Dimensões

5.3 – CONSIDERAÇÕES SOBRE A DEFINIÇÃO DE INDICADORES

Nas seções 5.2 e 5.3 foram definidos indicadores para a avaliação da qualidade de variáveis de fato e dimensões, respectivamente. Contudo, novos indicadores poderão ser definidos a qualquer momento, para avaliar outros tipos de objeto ou mesmo para aprimorar a avaliação da qualidade de variáveis de fato e dimensões.

A elaboração de um modelo para a qualidade, baseado em objetivos compostos por perspectiva, propósito, critério de qualidade, relacionados a tipos de objetos, tem justamente o intuito de conferir flexibilidade ao RMQ, para que ele possa ser utilizado em diversos tipos de situação. A definição de métricas, relacionadas a critérios de qualidade, para tipos de objeto do DW, funciona como um processo de instanciação do RMQ, que irá variar de acordo com o contexto e os propósitos da análise.

CAPÍTULO 6

EXEMPLO DE APLICAÇÃO: CENTRO DE PROCESSAMENTO DE DADOS DO ESTADO DO RIO DE JANEIRO (PRODERJ)

Neste capítulo será apresentado um exemplo de aplicação, que consiste na incorporação do AQUAWARE, proposto no capítulo 4, ao ambiente de *Data Warehouse* do PRODERJ, para fornecer informações sobre a qualidade dos dados para o módulo de consulta OLAP do Portal de Informações Consolidadas de Governo. Será apresentado ainda, um exemplo de consulta à qualidade de dados nesse ambiente, incluindo o detalhamento do comando de solicitação da consulta e a resposta retornada.

6.1 CENÁRIO: PRODERJ

O cenário do exemplo de aplicação é o Centro de Processamento de Dados do Estado do Rio de Janeiro (PRODERJ). O PRODERJ é uma autarquia vinculada à Secretaria de Ciência, Tecnologia e Inovação (SECTI), que tem como missão atuar como o órgão gestor de tecnologia da informação e comunicação, no âmbito do Governo do Estado do Rio de Janeiro.

O PRODERJ desempenha o importante papel de responsável pela proposição de diretrizes e orientações técnicas voltadas para o estabelecimento da política de tecnologia da informação e comunicação no âmbito do governo estadual. Dentre outras atribuições, ele atua como um grande birô de processamento de dados, integrando informações provenientes de diversos órgãos do governo estadual.

Dentre os projetos em andamento no PRODERJ atualmente, está o "Portal de Informações Consolidadas de Governo", que tem como objetivo apresentar para a alta administração do Governo do Estado, através de um portal na *Web*, informações gerenciais necessárias à tomada de decisão. Este projeto prevê a integração de diferentes bases de dados de órgãos do governo estadual, possibilitando o cruzamento de informações e a geração de *Data Marts* agrupados por área, como fazenda, segurança pública, educação, saúde etc, compondo um ambiente de *Data Warehouse* federado de

abrangência interorganizacional. É no contexto deste projeto que está inserido o exemplo de aplicação apresentado neste capítulo.

6.2 ARQUITETURA DO AMBIENTE DE DATA WAREHOUSE DO PRODERJ

No ambiente de *Data Warehouse* do PRODERJ, apresentado na figura 6.1, dados provenientes de DWs e sistemas transacionais, pertencentes a diferentes órgãos do governo estadual são utilizados para alimentar *Data Marts* organizados por assunto - saúde, segurança pública, fazenda, educação e outros - os quais são acessados individualmente ou de maneira integrada, compondo um *Data Warehouse* federado.

A arquitetura do DW federado é semelhante à proposta por Anand (1997), descrita na seção 3.1.1, sendo que o nível federado é composto por um *Data Mart* estratégico, que é modelado como uma visão materializada do esquema federado. O esquema federado, por sua vez, reúne informações consolidadas sobre diversas áreas do governo, obtidas dos diferentes *Data Marts* de assuntos. Esta abordagem foi escolhida por favorecer o aumento na velocidade de execução das consultas. Além disso, não existem restrições em relação ao processo de ETL, pois como a volatilidade dos dados não é muito alta, o intervalo de tempo entre duas cargas no *Data Mart* estratégico é suficiente para acomodar esse processo.

O *Data Mart* estratégico é utilizado pela alta administração do Governo do Estado do Rio de Janeiro, como fonte de informações gerenciais para tomadas de decisão. Dada à importância do papel desempenhado por este *Data Mart*, o AQUAWARE será acoplado à arquitetura do ambiente de *Data Warehouse* do PRODERJ, para que os tomadores de decisão possam ter conhecimento de informações sobre a qualidade dos dados utilizados nas análises, conferindo maior confiabilidade ao processo de tomada de decisão. A figura 6.1 ilustra a arquitetura do ambiente de *Data Warehouse* do PRODERJ, utilizando o AQUAWARE.

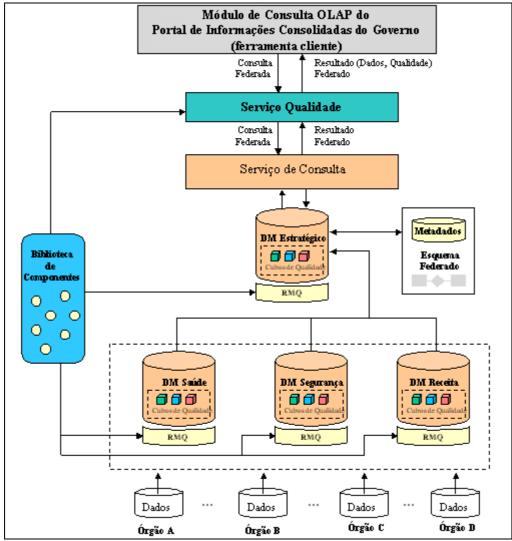


Figura 6.1 – Arquitetura do Ambiente de Data Warehouse do PRODERJ

Tanto os *Data Marts* de assunto quanto o *Data Mart* estratégico possuirão um repositório de metadados com informações sobre a qualidade de seus dados. O metamodelo (esquema) desse repositório será comum para todos eles, pois será utilizado o padrão definido para o AQUAWARE, apresentado na seção 4.5.3. A modelagem dos Cubos de Qualidade também será a mesma e corresponde à definida para o AQUAWARE, apresentada na seção 4.6. Seguindo a abordagem de Anand (1997), os Cubos de Qualidade estarão materializados também no *Data Mart* estratégico.

Quando o tomador de decisão, utilizando o módulo de consultas OLAP do "Portal de Informações Consolidadas do Governo" (ferramenta cliente) executa uma consulta federada, ele tem a possibilidade de solicitar também informações sobre a qualidade dos dados. A solicitação será enviada para o Serviço de Qualidade que, conforme descrito

na seção 4.3, irá enviar ao Serviço de Consulta uma consulta aos cubos de dados ou aos cubos de qualidade, retornando o resultado para a ferramenta cliente.

Vale ressaltar que a implementação do Serviço de Qualidade utilizando XMLA e a abordagem de Serviços *Web* facilita a integração do ambiente de *Data Warehouse* com o módulo de consulta OLAP do Portal de Informações Consolidadas de Governo, pois ambos utilizam XML para a troca de informações e a *Web* como infra-estrutura de comunicação.

6.3 OS CUBOS DE QUALIDADE

Os Cubos de Qualidade do PRODERJ são modelados segundo o padrão definido no AQUAWARE, descrito na seção 4.6, sendo que os graus de qualidade correspondem aos critérios identificados no capítulo 5, que são: acurácia sintática, acurácia semântica, completeza, atualidade e confiabilidade para fatos e dimensões e acurácia de conteúdo e precisão apenas para fatos. Como os indicadores de qualidade abrangem objetos do tipo "variável de fato" e "dimensão", o ambiente de *Data Warehouse* do PRODERJ possuirá os seguintes Cubos de Qualidade, apresentados nas figuras 6.2 a 6.5: Dimensões, Fatos, Anotações das Dimensões e Anotações do Fatos.

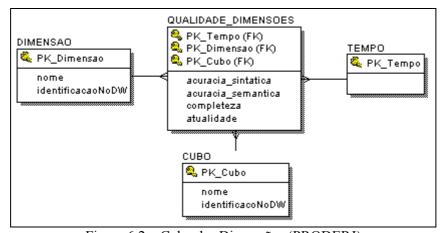


Figura 6.2 – Cubo das Dimensões (PRODERJ)

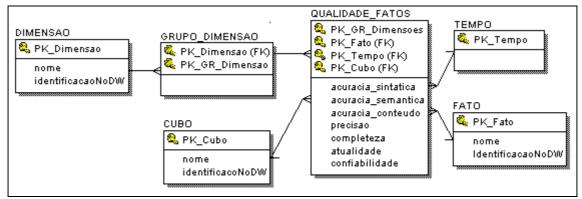


Figura 6.3 – Cubo dos Fatos (PRODERJ)

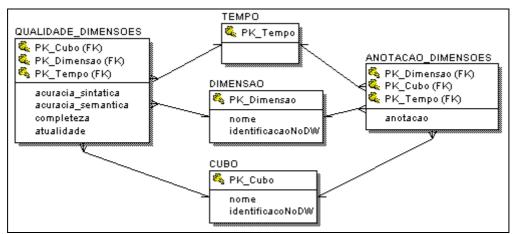


Figura 6.4 – Cubo de Anotações das Dimensões (PRODERJ)

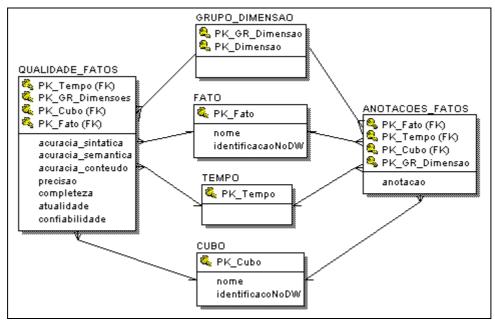


Figura 6.5 – Cubo de Anotações dos Fatos (PRODERJ)

6.4 DINÂMICA DE ATUAÇÃO DA BIBLIOTECA DE COMPONENTES

O processo que descreve a dinâmica de atuação dos componentes da Biblioteca de Componentes está ilustrado no Diagrama de Atividades da figura 6.6. Na figura, as raias representam os componentes responsáveis pela execução das atividades. O anexo III apresenta uma relação com os nomes de todos os componentes, acompanhados de sua descrição.

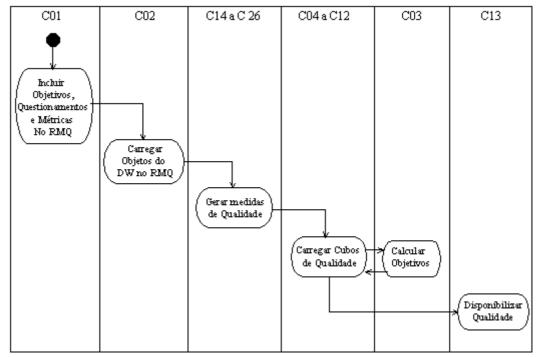


Figura 6.6 – Diagrama de Atividades da Dinâmica de Atuação dos Componentes

A primeira atividade que deve ser executada é a instanciação do metamodelo de qualidade com a inclusão dos objetivos, questionamentos e métricas, seguida da carga dos identificadores dos objetos do DW considerados. Serão utilizados os objetivos, questionamentos e métricas descritas no capítulo 5 para instanciar o metamodelo do RMQ no AQUAWARE no PRODERJ. A partir daí, sempre que for feita uma carga nos cubos de dados, devem ser carregados os objetos do DW que ainda não existam no RMQ e geradas as medidas de qualidade para aquela posição. A próxima etapa é a carga dos Cubos de Qualidade, que é feita em conjunto com o cálculo dos objetivos, que correspondem aos graus de qualidade. Concluída esta etapa, os dados de qualidade estarão disponíveis para serem consultados.

6.5 UMA CONSULTA À QUALIDADE DE DADOS NO *DATA MART* ESTRATÉGICO DO GOVERNO ESTADO DO RIO DE JANEIRO

Sendo o combate ao narcotráfico uma das prioridades do Plano de Governo do Estado do Rio de Janeiro, e buscando dar efetiva resposta ao aumento da criminalidade relacionada diretamente ao tráfico ilícito de entorpecentes, o Governo do Estado pretende implementar, ao longo do ano, ações voltadas ao combate a esta modalidade de crime. Assim, deverá identificar as áreas de maior ocorrência dos referidos delitos, bem como comportamentos diretamente relacionados com o uso de drogas, com o objetivo de estabelecer que localidades devem ser priorizadas quanto às ações a serem empreendidas. Assessorada pelo Secretário de Segurança, a Governadora do Estado recorre ao Portal de Informações Consolidadas de Governo em busca de informações que forneçam indícios sobre as localidades mais sensíveis. Por intermédio do Portal, ela tem acesso ao Data Mart estratégico, que contém informações provenientes de diferentes órgãos, interessantes para sua tomada de decisão, tais como: ocorrências de crimes relacionados ao tráfico e uso de entorpecentes (provenientes das delegacias civis), entradas de pacientes em hospitais apresentando problemas de saúde relacionados ao uso de drogas (provenientes dos hospitais) e relatos do serviço de orientação educacional das escolas sobre casos de alunos envolvidos com o consumo e porte de drogas (provenientes da Secretaria de Educação). Para ter maior segurança em relação à decisão que será tomada, é importante que o resultado das consultas seja acompanhado de informações sobre a qualidade dos dados apresentados.

A seguir é apresentado um exemplo de interação entre o Portal de Informações Consolidadas de Governo e o ambiente de *Data Warehouse* do PRODERJ, para a obtenção de informações de qualidade sobre uma das consultas realizadas pela Governadora, a saber:

"SELECIONAR

NÚMERO DE OCORRÊNCIAS DE CRIMES RELACIONADOS DIRETAMENTE AO NARCOTRÁFICO, NÚMERO DE ATENDIMENTOS EM HOSPITAIS RELACIONADOS AO USO DE ENTORPECENTES, QUANTIDADE DE ALUNOS ENVOLVIDOS COM DROGAS AGRUPADOS POR MUNICÍPIO DO ESTADO DO RIO DE JANEIRO

PARA O ANO DE 2003 "

Esta consulta é expressa na sintaxe MDX da seguinte forma:

Onde:

SEGURANCA_PUBLICA é um cubo do *Data Mart* estratégico.

QTD_CRIMES_ENTORPECENTES
QTD_PACIENTES_ENTORPECENTES
QTD_ALUNOS_ENTORPECENTES

MUNICIPIOS
TEMPO

é um cubo do *Data Mart* estratégico.

são variáveis de fato do cubo SEGURANCA_PUBLICA.

Considerando que a consulta em MDX seja submetida ao Serviço de Qualidade, via método *Execute* do XMLA-Q, a resposta retornada será conforme apresentado nas tabelas 6.1 e 6.2. Destaca-se que o formato de apresentação das informações de qualidade para o usuário final será definido pela ferramenta cliente, que neste exemplo é o módulo de consulta OLAP do Portal.

	Acurácia Sintática	Acurácia Semântica	Completeza	Atualidade	Anotação
MUNICIPIO	100	100	100	100	XML

Tabela 6.1 – Qualidade das Dimensões

Fatos	Acurácia	Acurácia	Acurácia	Precisão	Comple-	Atuali-	Confiabi-	Anotação
	Sintática	Semântica	de Conteúdo		teza	dade	lidade	
QTD_CRIMES_ ENTORPECENTES	100	100	100	100	100	100	100	XML
QTD_PACIENTES_ ENTORPECENTES	100	100	70	100	80	100	100	XML
QTD_ALUNOS_ ENTORPECENTES	100	100	100	80	80	80	100	XML

Tabela 6.2 – Qualidade dos Fatos em relação à dimensão MUNICIPIO

A descrição detalhada dos comandos em XMLA, envolvidos na presente interação entre o Portal e o Ambiente de *Data Warehouse*, é apresentada no anexo VI.

6.6 CONSIDERAÇÕES SOBRE A CONSULTA DE QUALIDADE

Sobre a consulta exemplificada na seção 6.5 pode-se concluir que tanto a dimensão MUNICIPIOS, quanto a variável de fatos QTD_CRIMES_ENTORPECENTES contêm informações bastante confiáveis, conforme evidenciado pelos graus de qualidade desses objetos. Já a variável de fatos QTD_PACIENTES_ENTORPECENTES deve ser analisada com cuidado, pois apresenta algumas restrições quanto à acurácia de conteúdo (os registros são feitos na emergência, podendo conter erros) e à completeza (não contempla todos os hospitais do RJ). O mesmo acontece com a variável de fatos QTD_ALUNOS_ENTORPECENTES que apresenta restrições quanto à precisão (os números são estimados pelos professores) e à completeza (não contempla todas as escolas do RJ).

CAPÍTULO 7

CONCLUSÕES E TRABALHOS FUTUROS

O objetivo desta dissertação foi definir as características de um ambiente computacional que, utilizado em conjunto com o *Data Warehouse*, forneça indicadores da qualidade dos dados para serem apresentados ao usuário final. A crescente utilização do *Data Warehouse* para apoiar processos de tomada de decisão tem provocado um aumento da preocupação em relação à qualidade dos mesmos. Dados de má qualidade custam dinheiro e reduzem a produtividade — o tempo gasto com o diagnóstico e solução de problemas poderia ser utilizado de forma produtiva. A possibilidade de tirar vantagem dos recursos de informação só é possível se existe um conhecimento prévio do quão confiáveis eles são. Em contextos interorganizacinais a necessidade de tornar disponíveis informações de qualidade, é ainda maior, pois como grande parte das fontes de dados são externas à organização, o usuário não tem como prever o grau de qualidade dos mesmos.

Para atender à necessidade de interoperabilidade em ambientes federados, autônomos e heterogêneos, característicos do contexto interorganizacional, foi utilizada a abordagem de serviços *Web* acrescida do uso do padrão XMLA, para a implementação do Serviço de Qualidade. Para armazenar os graus de qualidade fornecidos por esse serviço foram criados Cubos de Qualidade, os quais são alimentados periodicamente a partir das medidas de qualidade existentes no RMQ. A elaboração do Metamodelo do RMQ como uma extensão do CWM possibilita a integração do AQUAWARE com qualquer ferramenta cujo repositório de metadados esteja modelado segundo esse padrão. A utilização do mecanismo de indexação global única de Campos (1993) permite que qualquer objeto do DW seja referenciado no metamodelo do repositório, independente do nível da arquitetura multi-camadas ao qual pertença. A representação da qualidade no metamodelo segundo o paradigma GQM orienta a obtenção de objetivos questionamentos e métricas, através de uma abordagem do tipo *top-down*.

A principal contribuição desse trabalho é a viabilização do fornecimento de informações sobre a qualidade dos dados do DW para ferramentas de consultas analíticas, para que as mesmas possam formatá-las e apresentá-las para o usuário final.

O fornecimento de informações sobre a qualidade de dados para os tomadores de decisão representa um avanço em relação às consultas analíticas convencionais, pois possibilita que ele avalie até que ponto pode confiar nos dados apresentados, conferindo mais confiabilidade ao processo de tomada de decisão.

Dentre outras contribuições estão a definição do XMLA-Q como uma evolução do XMLA para contemplar funcionalidades relativas à qualidade; a definição de indicadores para a avalição da qualidade de dados, sob a forma de objetivos, questionamentos e métricas inspiradas nos critérios voltados para o usuário final, definidos por Wang et al. (1996) e Wand e Wang (1996), e nos critérios caracaterísticos de contextos interorganizacionais, definidos por Bertolazzi e Scanapieco (2001) e Aubert et al. (2003); e a incorporação do AQUAWARE ao Projeto do Portal de Informações Consolidadas de Governo, no PRODERJ, contribuindo para a evolução dos processos decisórios do Governo do Estado do Rio de Janeiro. Além disso, foram comparadas diversas abordagens existentes para a avaliação da qualidade no DW e realizado um levantamento sobre critérios de qualidade, ainda sem um consenso na comunidade científica sobre que conjunto de critérios melhor expressa o conceito de qualidade de dados.

Este trabalho apresenta algumas limitações e deixa alguns problemas em aberto, que poderão ser resolvidos em trabalhos futuros. A seguir são apresentadas algumas sugestões:

- Desenvolvimento de um mecanismo que utilize as informações de qualidade para propor melhorias nos processos relativos à informação, reduzindo assim, os custos ocasionados pela má qualidade;
- Implementação das funcionalidades necessárias para a interação com o AQUAWARE numa ferramenta de consulta analítica, habilitando-a a receber as informações de qualidade e apresentá-las para o usuário final, junto com os resultados das consultas;
- Definição de indicadores de qualidade para outros objetos do DW, além dos fatos e das dimensões, e a especificação de componentes para a medição da qualidade dos mesmos;

Definição de indicadores de qualidade sob diferentes perspectivas, dentre elas a
do administrador do DW, de forma que os resultados da avaliação da qualidade
possam ser utilizados como subsídios para o aprimoramento do processo de
administração do ambiente de *Data Warehouse*.

Além destes, já está em andamento a implementação de um protótipo do AQUAWARE que será utilizado no ambiente de *Data Warehouse* do PRODERJ para, num primeiro momento, prover informações de qualidade para o Portal de Informações Consolidadas do Governo.

BIBLIOGRAFIA

ABATE, M. L. et al. *A Hierarquical Approach to Improving Data Quality*. Data Quality Journal, USA, v.4, n.1, 1998. Disponível em: http://www.dataquality.com/998abate.htm Acesso em: 08 Jun 2003.

ABREU M.: entrevistas [Jul a Nov 2003]. Entrevistadores: Glenda Amaral e Maria Luiza Campos. Rio de Janeiro: PRODERJ, 2003. Entrevistas sobre o problema da qualidade de dados em *Data Warehouse* no PRODERJ e o Projeto do Portal de Informações Consolidadas de Governo.

ANAND, V.J. *A Data Warehouse within a Federated Database Architecture*. Association for Information Systems 1997 Americas Conference, Indianapolis, Indiana, USA, 1997. Disponível em: http://hsb.baylor.edu/ramsower/ais.ac.97/papers/anand.htm. Acesso em: 04 Nov 2003.

AUBERT, A.A.; VANDENBOSH, B.; MIGNERAT, M. *Towards the Measurement of Proocess Integration*, Scientific Series 2003s-04, Montreal, Canadá, 2003. Disponível em: http://www.cirano.qc.ca/pdf/publication/2003s-4.pdf. Acesso em: 24 Out 2003.

BAHIA, A. DE SÁ. *Avaliação da Complexidade de Software Científico*. 1992. Dissertação de Mestrado – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

BALLOU, D.P.; PAZER, H.L. *Modeling Data and Process Quality in Multi-input, Multi-output Information Systems*, Management Science, v.31, n.2, pp.150-162, 1985.

BASILI, V.R.; WEISS, D.M. *A metodology for colecting valid software engineering data*. IEEE Transactions on Software Engineering, v.10, n.6, pp.728-738, 1984.

BASILI, V.R.; ROMBACH, H.D. *The TAME Project Towards Improvement-Oriented Software Environments*, IEEE Transactions on Software Engineering, v.14, n.6, 1988. Disponível em: http://ieeexplore.ieee.org/xpl/abs_free.jsp?arNumber=6156. Acesso em: 15 Nov 2003.

BASILI, V.R.; CALDIERA, G.; ROMBACH, H.D. *The Goal Question Metric Approach*. Encyclopedia of Software Engineering, pp.528-532, John Wiley & Sons, 1994. Disponível em: http://www.cs.umd.edu/users/basili/papers.html. Acesso em: 11 Nov 2003.

BELCHIOR, A.D. *Controle da Qualidade de Software Financeiro*. 1992. 183p. Dissertação de Mestrado – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

BERTOLAZZI, P.; SCANNAPIECO, M. *Introducing Data Quality in a Cooperative Context*. In: Proceedings of the 6th International Conference on Information Quality, Boston, MA, USA, 2001. Disponível em:

http://www.dis.uniroma1.it/~monscan/ResearchActivity/Publication.htm. Acesso em: 15 Nov 2003.

BLASCHEK, J.R.S; ROCHA, A.R.C.DA. *Requisitos de Qualidade de Sistemas de Informação*. Publicações Técnicas, Núcleo de Pesquisa e Desenvolvimento de Software Médico, Fundação Baiana de Cardiologia, n.4, 1995.

BOBROWSKI, M.; MARRÉ, M.; YANKELEICH, D. *Measuring Data Quality*. 1999. Relatório Técnico – Departamento de Computação – Faculdade de Ciências Exatas e Naturais, Universidade de Buenos Aires, Buenos Aires. Disponível em: http://www.dc.uba.ar/people/proyinv/1999/resumen.html. Acesso em: 15 Nov 2003.

CAMPOS, M.L.M. *A Data Modeling Architecture for Integrated Schema and Data*. 1993. 144p. Tese de Doutorado, University of East Anglia, Inglaterra.

CAMPOS, F.C.A. *Hipermídia na Educação: Paradigmas e Avaliação da Qualidade*. 1994a. 137p. Dissertação de Mestrado – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

CAMPOS, G.H.B. *Metodologias para avaliação da qualidade de software educacional: Diretrizes para desenvolvedores e usuários*. 1994b. 232p. Tese de Doutorado –

COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

CAMPOS, M.L.A. A Organização de Unidades do Conhecimento em Hiperdocumentos: o modelo conceitual como um espaço comunicacional para a realização da autoria. 2001. 198p. Tese de Doutorado – Universidade Federal do Rio de Janeiro, Escola de Comunicações, Instituto Brasileiro de Informação em Ciência e Tecnologia, Rio de Janeiro. Disponível em: http://genesis.nce.ufrj.br/dataware. Acesso em: 01 Dez 2003.

CAMPOS, F.: entrevista [Out 2003]. Entrevistadora: Glenda Amaral. Rio de Janeiro: NCE, 2003. Entrevista sobre experiências em projetos de *Data Warehouse*.

CAVALCANTI, M.; GOMES, E. *A Nova Riqueza das Organizações: Os Capitais do Conhecimento*, Revista Brasileira de Tecnologia e Negócios de Petróleo, Petroquímica, Química Fina, Gás e Indústria do Plástico, Ano II, n. 16, 2000. Disponível em: http://www.crie.coppe.ufrj.br. Acesso em: 30 Nov 2003.

CARVALHO, D.O. *Qualidade de Sistemas de Informação Hospitalar*. 1997. 229p. Dissertação (Mestrado) – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

COMERLATO, C.; XEXEO, G.B.; ROCHA, A.R.C.DA. *Avaliação da reutilizabilidade de Componentes de Software*. In: VIII Simpósio Brasileiro de Engenharia de Software, Curutiba, Paraná, 1994, pp.23-26.

CYKANA, P; PAUL,A.;STERN,M. *DoD Guidelines on Data Quality Management*. In: Proceedings of the 1996 Conference on Information Quality. Cambridge, MA, 1996, pp. 154-191. Disponível em: http://www-datadmn.itsi.disa.mil/dqpaper.pdf. Acesso em: 10 Jun 2003.

DELONE, W.H., MCLEAN, E.R. *Information System Success: The Quest for the Dependent Variable*. Information Systems Research, v.3, n.1, 1992, pp 60-95.

ECKERSON, W. W. *Data Quality and The Botttom Line: Achieving Business Success through a Commitment to High Quality Data*. The Data Warehousing Institute Report Series, USA. Disponível em: http://www.stuart.iit.edu/courses/im510/spring2002/dqreport.pdf. Acesso em: 11 Dez 2002.

ELIAS. M.; MARTINHO, L.: entrevista [Set 2003]. Entrevistadores: Glenda Amaral e Maria Luiza Campos. Rio de Janeiro: Hiper Consultoria, 2003. Entrevista sobre experiências em projetos de *Data Warehouse*.

ENGLISH, L.P. Improving Data Warehouse and Business Information Quality – Methods for Reducing Costs and Increasing Profits. USA: Wiley, 1999. 518p.

EVERETT, D. *Compare and Contrast JOLAP and XML for Analysis*, Hyperion Developer Network, 2003. Disponível em:

http://dev.hyperion.com/resource_library/articles/jolap_xmla.cfm. Acesso em: 30 Nov 2003.

FERREIRA, A.B.H *Aurélio Século XXI: O dicionário da língua portuguesa*. 3ed. Rio de Janeiro: Nova Fronteira, 1999.

GARDYN, E. *A Data Quality Handbook for a Data Warehouse*. In: Proceedings of the 1997 Conference on Information Quality. Cambridge, MA, 1997, pp.267-290.

GOODHUE, D.L., *Understanding User Evaluations of Information Systems*, Management Science, v.41, n.12, 1995, pp. 1844-1927.

HUBER, J. et al. *Integrating Data Warehouses versus Building a Federated Data Warehouse: A Comparision*. In: Proceedings of the 3rd International Conf. on Data Warehousing and Knowledge Discovery, Munich, Germany, 2001. Disponível em: http://www.scch.at/servlet/resource.ResourceLoader?id=120. Acesso em: 04 Nov 2003.

HYPERION SOLUTIONS, SUN MICROSYSTEMS. *Java TM OLAP Interface* (*JOLAP*), Proposed Final Draft, 2003. Disponível em:

http://jcp.org/aboutJava/communityprocess/first/jsr069/index.html. Acesso em: 04 Nov 2003.

INMON, W. H. *Building the Corporate Information Factory from a Blueprint – Part I*, 2001a. Disponível em: http://www.billinmon.com. Acesso em: 11 Jun 2003.

Building the Corporate Information Factory from a Blueprint – Pa	art II,
2001b. Disponível em: http://www.billinmon.com. Acesso em: 11 Jun 2003.	

Building the Corporate Information Factory from a Blueprint – Part III,
2001c. Disponível em: http://www.billinmon.com. Acesso em: 11 Jun 2003.
Enterprise Intelligence – Enabling High Quality In the Data Warehouse
DSS Environment. Disponível em: http://www.ascentialsoftware.com/cgi-
bin/dataquality.cgi?URL=WP_Inmon_3.pdf. Acesso em: 10 Jun 2003.
ISO 8402, Quality Vocabulary, ISO DIS.1994.
ISO 9126, Information Tecnology – Software Product Evaluation – Quality
Characteristics and Guidelines for their use, ISO/IEC. 1991.
JARKE, M.; VASSILIOU, Y. Data Warehouse Quality: A Review of the DWQ Project.
In: Proceedings of the 1997 Conference on Information Quality. Cambridge, MA, pp.
229-313, 1997. Disponível em: http://www.dblab.ntua.gr/~dwq/iq97_dwq.pdf. Acesso
em: 08 Jun 2003.
JEUSFELD, M. A.; QUIX, C.; JARKE, M. Design and Analysis of Quality Information
for Data Warehouses, 1998. Disponível em:
$http://www.dbnet.ece.ntua.gr/{\sim}dwq/p48.pdf \ . \ Acesso\ em:\ 11\ Nov\ 2003.$
KIMBALL, R. et al. The Data Warehouse Lifecycle Toolkit – Expert Methods for
Designing, Developing and Deployind Data Warehouses. New York: John Wiley &
Sons, Inc., 1998. 771p.
<i>Indicators of Quality</i> . Intelligent Enterprise, v.3, n.6, 2000. Disponível
em: http://www.intelligententerprise.com/000410/webhouse.shtml. Acesso em: 10 Nov
2003.
KOVAK, R.; LEE, Y.W.; PIPINO, L.L. Total Data Quality Management: The Case of
IRI. In: Proceedings of the 1997 Conference on Information Quality, Cambridge, MA,
1997, pp.63-79. Disponível em:
http://web.mit.edu/tdqm/www/tdqmpub/IRITDQMCaseOct97.pdf. Acesso em: 09 Jun

2003.

LAUDON. K.C. Data Quality and due process in large interorganizational record systems. 1986. Commun, ACM 29, 1, pp.4-11.

LEE. Y. W. et al. AIMQ: *A Methodology for Information Quality Assessment*, Information & Management, v.40, n.2, 2001, pp.133-146. Disponível em: http://web.mit.edu/tdgm/www/tdgmpub/AIMQJun02.pdf. Acesso em: 08 Jun 2003.

LOSHIN, D. *Enterprise Knowledge Management – The Data Quality Approach*. USA: Morgan Kaufman, 2001. 493p.

MANDKE, V.V.; NAYAR, M.K. *Information Integrity – A Structure for its Definition*. In: Proceedings of the 1997 Conference on Information Quality, Cambridge, MA, 1997, pp.314-338.

MANGISENGI, O.; HUBER, J. *A Framework for Supporting Interoperability of Data Warehouse Islands using XML*. In: Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery, Minich, Germany, 2001, pp.328-338. Disponível em: http://www.scch.at/servlet/resource.ResourceLoader?id=217. Acesso em: 04 Nov 2003.

MARCHAND, D. *Managing information quality*. In: WORMELL, I. (Ed.) Information quaity: definitions and dimensions. London: Taylor Graham, 1990. p. 7-17.

MARCO, D. Meta Data and Data Administration: Implementing Data Quality Through Meta Data, Part 1. DM Review, 2000a. Disponível em: http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=2104. Acesso em: 10 Jun 2003.

_____. Meta Data and Data Administration: Implementing Data Quality Through Meta Data, Part 2. DM Review, 2000b. Disponível em:

http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=2176. Acesso em: 10 Jun 2003.

MATSUMURA, A.; SHOURABOURA, N. *Competing with Quality Information*. In: Proceedings of the 1996 Conference on Information Quality, Cambridge, MA, 1996, pp.72-86.

MEYEN, D.M.; WILLSHIRE, M.J. *A Data Quality Engineering Framework*. In : Proceedings of the 1997 Conference on Information Quality. Cambridge, MA, 1997, pp.95-116.

MICROSOFT CORPORATION, HYPERION SOLUTIONS CORPORATION. *XML for Anaysis Specification*, v.1.1, 2002. Disponível em: http://www.xmla.org/docs_pub.asp. Acesso em: 04 Nov 2003.

MICROSOFT CORPORATION. *Microsoft OLE DB - OLE DB Programmer's Reference - Part 3: OLE DB for On-line Analytical Processing (OLAP)*. Disponível em: http://msdn.microsoft.com/library/default.asp?url=/library/en-us/oledb/htm/olprpart3.asp; Acesso em: 07 de novembro de 2003.

NAUMANN, F. *Data Fusion and Data Quality*. In: Proceedings of the New Techniques and Technologies for Statistics Seminar. Sorrent, Itália, 1998. Disponível em: http://citeseer.nj.nec.com/112371.html. Acesso em: 15 Nov 2003.

NAUMANN, F.; LESER, U. *Quality-driven Integration of Heterogeneous Information Systems*. In: Proceeding of the 25th International Conference on Very Large Databases. Edinburg, UK, 1999, pp. 447-458. Disponível em: http://citeseer.nj.nec.com/naumann99qualitydriven.html. Acesso em: 15 Nov 2003.

NAUMANN, F.; ROLKER, C. *Assessment Methods for Information Quality Criteria*. In: Proceedings of the Conference on International Quality, Boston, 2000. Disponível em: : http://citeseer.nj.nec.com/295186.html. Acesso em: 15 Nov 2003.

NEHMY, R.M.Q; PAIM, I. *A desconstrução do conceito de "qualidade da informação*", Revista Ciência da Informação, v.27, n.1, p.36-45, 1998. Disponível em: http://www.ibict.br/cionline/270198/27019805.htm. Acesso em: 11 Mar 2003.

OIVO, M.; BASILI, V. *Representing Software engineering models: the TAME goal-oriented approach*. IEEE Transactions on Software Engineering, v.18, n.10, pp-886-898, 1992. Disponível em: http://www.cs.umd.edu/users/basili/papers.html. Acesso em: 11 Nov 2003.

OLAISEN, J. *Information quality factors and the cognitive authority of eletronic for the cognitive authority of eletronic information*. In: WORMELL, I. (Ed.) Information quality: definitions and dimensions. London: Taylor Graham, 1990. p. 84-91.

OLIVEIRA, K. N., *Avaliação da Qualidade de Sistemas Especialistas*. 1995. 170p. Dissertação de Mestrado – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

OLLE, T.W. *Data Levels in IRDS, The Future of Data Dictionaries*, The British Computer Society Data Management Specialist Group, Simon Holloway (ed.), 1988, pp.31-48.

OMG (OBJECT MANAGEMENT GROUP). Common Warehouse Metamodel Specification. 2003. Disponível em: http://www.omg.org. Acesso em: 17 Ago 2003.

PALERMO. S.; ROCHA, A.R.C.DA. An experience on Evaluating Software Quality for High energy Physics. Computer Physics Communications, 1989.

PEREIRA JUNIOR, O.A. *Qualidade de Dados em Sistemas de Informação Ambientais*. 2000. 92p. Dissertação de Mestrado – Universidade Federal do Rio de Janeiro, Escola de Comunicações, Instituto Brasileiro de Informação em Ciência e Tecnologia, Rio de Janeiro.

PINHO, S.F.C. *Avaliação da Qualidade de Dados pela Não Conformidade*. 2001. 110p. Dissertação de Mestrado – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

PIPINO. L.L. et al. *Data Quality Assessment*. Communications of the ACM, USA, v.45, n.4, 2002. Disponível em:

http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf. Acesso em: 11 Fev 2003.

REDMAN, T.C. *Data Quality: Management and Tecnology*. Bantam Books: New York, NY, 1992. 308p.

REPO, A.J. The value of information: approaches in economics, accounting and management science. Journal of American Science, v. 40, n. 2, p. 68-85, 1989.

ROCHA, A.R.C.DA. *Um modelo para avaliação da qualidade de especificações*. 1983. Tese de Doutorado – PUC-Rio, Rio de Janeiro.

ROTHENBERG, J. *Metadata to Support Data Quality and Longevity*. 1996. Disponível em: http://www.computer.org/conferences/meta96/rothenberg_Paper/ieee.data-quality.html. Acesso em: 13 Jan 2003.

SPOFFORD, G. *Access to Intelligence: The New OLAP API*, Intelligent Enterprise, 2002. Disponível em: http://www.intelligententerprise.com/020917/515feat2_2.shtml. Acesso em: 30 Nov 2003.

STRONG, D.M; LEE, Y.W.; WANG, R.Y. *Data Quality in Context*, Communications of The ACM, v.40, n.5, 1997. Disponível em: http://web.mit.edu/tdqm/www/tdqmpub/StrongLeeWangCACMMay97.pdf. Acesso em: 08 Jun 2003.

TSALGATIDOU, A.; PILIOURA, T. *An Overview of Standards and Related Technology in Web Services*, International Journal of Distributed and Parallel Databases, Special Issue on E-Services, v.12, n.2, pp.135-162, 2002. Disponível em: http://cgi.di.uoa.gr/~afrodite/publications.html. Acesso em 30/11/2003.

WAND, Y.; WANG, R.Y. *Anchoring Data Quality Dimensions in Ontological Foundations*, Communications of the ACM, v.39, n.11, 1996, pp.86-95. Disponível em: http://web.mit.edu/tdqm/www/tdqmpub/WandWangCACMNov96.pdf. Acesso em: 08 Jun 2003.

WANG, R.; STOREY, V.; FIRTH, C. *A Framework for Analysis of Data Quality Research*, IEEE Transactions on Knowledge and Data Engineering, v.7, n.4, 1995, pp. 623-640. Disponível em:

http://web.mit.edu/tdqm/www/tdqmpub/SURVEYIEEEKDEAug95.pdf. Acesso em: 08 JUN 2003.

WANG, R.; STRONG, D.; GUARASCIO, L.M. *Beyond Accuracy: What Data Quality Means to Data Consumers*, Journal of Management Information Systems, v.12, n.4, pp. 5-33, 1996. Disponível em: http://web.mit.edu/tdqm/www/papers/94/94-10.html. Acesso em: 08 Jun 2003.

VASSILIADIS, P. *Data Warehouse Modeling and Quality Issues*. 2000. 129p. Tese de Doutorado - National Technical University of Athens, Grécia. Disponível em: http://citeseer.nj.nec.com/vassiliadis00data.html. Acesso em: 15 Nov 2003.

ZHU, Y; BUCHMANN, A. *Evaluating and Selecting Web Sources as External Information Resources of Data Warehouse*. In Proceedings of the 3rd International Conference on Web Information Systems Engineering, Singapore, 2002. Disponível em: http://csdl.computer.org/comp/proceedings/wise/2002/1766/00/17660149abs.htm. Acesso em: 15 Nov 2003.

ZMUD, R. Concepts, Theories and Techniques: An Empirical Investigation of the Dimensionality of the Concept of Information. Decision Sciences, v.9, n.2, 1978, pp.187-195.

ANEXO I

DESCRIÇÃO DOS MÉTODOS DISCOVER E EXECUTE DO XMLA

MÉTODO DISCOVER

Usado para recuperar metadados do ambiente OLAP.

Sintaxe:

```
Discover (
    [in] RequestType As EnumString,
    [in] Restrictions As Restrictions,
    [in] Properties As Properties,
    [out] Result As Rowset)
```

Parâmetros:

RequestType [in]

Especifica o tipo de informação que deve ser retornada. Corresponde a um valor pertencente à enumeração definida para o *Request Type*. Cada *Request Type* está associado a um *rowset* do tipo de objetos *Rowset* do XMLA (MICROSOFT, HYPERION, 2002). Os *rowsets* correspondem aos formatos no qual as informações para cada tipo de *Request Type* serão retornadas. Assim, existe um *rowset* para cada valor enumerado do parâmetro *Request Type*. Caso o servidor defina novos valores para esse parâmetros, devem ser definidos também os *rowsets* correspondentes. A especificação do XMLA (MICROSOFT, HYPERION, 2002) contém a descrição do tipo de objetos *Rowset*.

Valor enumerado	Descrição
DISCOVER_DATASOURCES	Retorna a lista de fontes de dados disponíveis no servidor, que podem ser acessadas via XMLA.
DISCOVER_PROPERTIES	Retorna uma lista de informações e valores sobre todas as propriedades que são suportadas pelo servidor.
DISCOVER_SCHEMA_ROWSETS	Retorna os nomes, valores e demais informações sobre todos os tipos de <i>Request Type</i> .
DISCOVER_ENUMERATORS	Retorna a lista de nomes, tipos de dados e valores enumerados de todas as enumerações suportadas pelo servidor.
DISCOVER_KEYWORDS	Retorna um rowset contendo a lista de palavras-chave

reservadas para o servidor.

DISCOVER_LITERALS Retorna informações sobre os literais suportados pelo

servidor.

Constante de Esquema Rowset Dada uma constante que corresponde a um rowset

definido no OLE DB, retorna esse mesmo *rowset* em XML. Os servidores podem também especificar *rowsets*

próprios.

Restrictions [in]

Permite a aplicação de restrições sobre os resultados retornados em *Result*. As colunas de *Result* são definidas pelo *rowset* especificado no parâmetro *RequestType*. Algumas colunas de *Result* podem ser especificadas para filtrar o resultado retornado. Essas colunas e as restrições possíveis estão descritas junto com os *rowsets* do XMLA. Para obter informações sobre restrições usando *rowsets* específicos do servidor, o parâmetro DISCOVER SCHEMA ROWSETS deve ser usado em *Request Type*.

Properties [in]

Este parâmetro é do tipo *Properties* do XMLA (MICROSOFT, HYPERION, 2002). Consiste de uma coleção de propriedades do XMLA. Cada propriedade permite que o usuário controle alguns aspectos do método *Discover*, como por exemplo, especificar o formato do resultado retornado. As propriedades disponíveis podem ser obtidas através do *Request Type* DISCOVER_PROPERTIES. A especificação do XMLA contém a descrição detalhada do tipo *Properties* (MICROSOFT, HYPERION, 2002).

Result [out]

Contém o resultado retornado pelo servidor, como um objeto do tipo Rowset. As colunas e o conteúdo do resultado são especificados em *Request Type* e em *Restrictions*. O formato do resultado também é determinado pelos valores especificados em *Request Type*.

MÉTODO EXECUTE

Usado para requisitar uma ação ao servidor: consulta a dados e atualização de dados.

Sintaxe:

```
Execute (
   [in] Command As Command,
   [in] Properties As Properties,
   [out] Result As Resultset)
```

Parâmetros

Command [in]

Sentença em MDX, encapsulada num comando mdXML ou comandos específicos do servidor.

Properties [in]

Este parâmetro é do tipo *Properties* do XMLA (MICROSOFT, HYPERION, 2002). Consiste de uma coleção de propriedades do XMLA. Cada propriedade permite que o usuário controle alguns aspectos do método *Execute*, tais como definir as informações necessárias para a conexão, especificar o formato do resultado retornado, etc. As propriedades disponíveis podem ser obtidas através do *Request Type* DISCOVER_PROPERTIES. A especificação do XMLA contém a descrição detalhada do tipo *Properties* (MICROSOFT, HYPERION, 2002).

Result [out]

Este parâmetro é do tipo *Resultset* e contém o resultado retornado pelo servidor. O formato do *Resultset* é definido a partir dos parâmetros *Command* e *Properties*. A especificação do XMLA contém a descrição detalhada do tipo *Properties* (MICROSOFT, HYPERION, 2002).

ANEXO II

DESCRIÇÃO DOS CRITÉRIOS DE QUALIDADE DA TABELA 2.5

ACURÁCIA: representa a distância entre dois valores v e v', sendo v considerado o valor correto.

COMPLETEZA: é o grau no qual os elementos de um esquema estão presentes nas instâncias.

"TEMPORALIDADE": disponibilidade dos dados no tempo esperado, de acordo com os requisitos de tempo especificados pelo destino.

CONSISTÊNCIA: grau no qual dados pertencentes a diferentes conjuntos estão consistentes entre si.

RELEVÂNCIA: grau no qual os dados são aplicáveis.

ACESSIBILIDADE: grau de disponibilidade dos dados, facilidade e rapidez com a qual eles podem ser recuperados.

CONFIABILIDADE: expressa o grau de confiança que usuários têm nos dados

ATUALIDADE: é a diferença de tempo entre o instante em que o dado é atualizado na origem e o instante em que ele é disponibilizado.

PRECISÃO: grau no qual o valor do dado corresponde a um valor aproximado em relação ao valor real.

QUANTIDADE APROPRIADA: quantidade de dados é adequada ap propósito em questão.

FACILIDADE DE ENTENDIMENTO: grau no qual os dados estão descritos em linguagem e unidades adequadas e definidos de forma clara.

IMPORTÂNCIA: importância dos dados para o destino. Está relacionada com a possibilidade de existirem processos no destino que dependam dos dados.

DETALHAMENTO ADEQUADO: a granularidade do dado está adequada às necessidades.

REPRESENTAÇÃO CONSISTENTE: o dado é apresentado de forma consistente com o seu domínio e com os demais dados similares.

UNICIDADE: cada objeto de dado é identificado de forma única.

SEGURANÇA: existência de mecanismos que garantam o controle de acesso aos dados.

USABILIDADE: facilidade de uso.

UTILIDADE: utilidade dos dados para o propósito em questão.

PRIVACIDADE: existência de mecanismos que garantam a privacidade de acesso aos dados (graus de sigilo, por exemplo).

REPUTAÇÃO: corresponde à credibilidade da organização de origem, que atua como fonte dos dados.

OBJETIVIDADE: imparcialidade no sentido que a sua interpretação é única e independe do ponto de vista do observador.

VALIDADE: grau no qual o conteúdo do dado é válido.

VALOR AGREGADO: grau dos benefícios proporcionados e vantagens do uso.

CONCISÃO: os dados são representados de forma compacta e não são prolixos.

NATURALIDADE: refere-se ao relacionamento entre o objeto modelado e o mundo real, significando que cada atributo representa um fato sobre um objeto do mundo real e os valores do domínio ao qual pertence espelham corretamente realidade.

HOMOGENEIDADE: atributos possuem apenas um significado, ou seja, não existe mais de um tipo de informação sendo representado pelo mesmo atributo.

CONTROLE DE REDUNDÂNCIA: as redundâncias existentes são apenas aquelas planejadas com propósitos específicos.

CONSISTÊNCIA ESTRUTURAL: atributos similares são modelados de forma consistente, pertencendo ao mesmo domínio.

PORTABILIDADE: a interface de apresentação dos dados suporta a migração da aplicação para plataformas diferentes.

FLEXIBILIDADE: a modelagem dos dados é capaz de refletir mudanças para atender a novos requisitos, sem que a mesma precise ser alterada.

TRATAMENTO DE VALORES NULOS: possibilidade de identificar o motivo pelo qual os dados não estão presentes: se estão indisponíveis, se não aplicáveis, se são desconhecidos ou se não existe valor no domínio que os represente corretamente.

ARMAZENAMENTO EFICIENTE: há uma gerência sobre o espaço ocupado em disco.

FACILIDADE DE OBTENÇÃO: facilidade com que os dados são coletados e armazenados.

ROBUSTEZ: habilidade de refletir mudanças do mundo real no modelo de dados, alterando-o o mínimo possível.

AUSÊNCIA DE AMBIGÜIDADES: cada item de informação tem um único significado.

REGISTROS DE APELIDOS: apelidos (*alias*) dos dados estão associados aos mesmos.

CONTROLE DE VERSÕES: controle das diferentes versões dos dados.

ORIGEM IDENTIFICADA: explicitação da origem dos dados.

REPRESENTAÇÃO APROPRIADA: o formato de apresentação dos dados está adequado aos propósitos.

CONCORDÂNCIA SOBRE O USO: grau no qual o dado atende às necessidades de grupos diferentes de usuários de forma que eles concordem a utilizá-lo ao invés de criar elementos próprios.

INTEGRIDADE: manutenção consistente das estruturas de dados e do relacionamento entre entidades.

EXISTÊNCIA DE METADADOS: existem descrições sobre o significado dos dados.

SIMPLICIDADE: o modelo de dados não é complexo, sendo de fácil compreensão.

ACOMPANHAMENTO: existe um responsável pelo dado.

UBIQÜIDADE: grau no qual diferentes departamentos de empresa compartilham o dado.

CUSTO ADEQUADO: os custos de obtenção, gerência da qualidade, armazenamento e outros custos relacionados ao dado estão dentro do valor previsto.

ANEXO III

LISTA DOS COMPONENTES DA BIBLIOTECA

Componente	Nome			
C01	Inclusão dos Objetivos Questionamentos e Métricas			
C02	Carga dos Objetos do DW no RMQ			
C03	Cálculo do Objetivo			
C04	Carga da dimensão DIMENSAO			
C05	Carga da diemensão TEMPO			
C06	Carga da tabela de fatos do Cubo das Dimensões			
C07	Carga da dimensão GRUPO_DIMENSAO			
C08	Carga da dimensão FATO			
C09	Carga da tabelas de fatos do Cubo dos Fatos			
C10	Carga da dimensão CUBO			
C11	Carga da tabela de auditoria das tuplas			
C12	Carga da tabela de auditoria das células			
C13	Serviço de Qualidade			
C14	Geração das medidas de qualidade para as métricas M1.1.1 e M10.1.1 (Registros que obedecem às regras de formação sintática)			
C15	Geração das medidas de qualidade para as métricas M1.2.1 e M10.2.1 (Registros que pertencem ao domínio)			
C16	Geração das medidas de qualidade para as métricas M1.3.1 e M10.3.1 (Registros que pertencem ao tipo de dados)			
C17	Geração das medidas de qualidade para as métricas M2.1.1 e M11.1.1 (Registros que obedecem às regras de negócio)			
C18	Geração das medidas de qualidade para a métrica M2.2.1 (Registros contendo chaves estrangeiras cadastradas nas dimensões)			
C19	Geração das medidas de qualidade para a métrica M3.1.1 (Registros que estão em conformidade com fonte conhecida)			
C20	Geração das medidas de qualidade para a métrica M4.1.1 (Proximidade do valor armazenado em relação ao valor real).			
C21	Geração das medidas de qualidade para a métrica M5.1.1 (Registros onde o campo está preenchido ou está identificado como não aplicável)			
C22	Geração das medidas de qualidade para a métrica M5.2.1 e M12.1.1 (Membros existentes que estão cadastrados nas dimensões)			
C23	Geração das medidas de qualidade para a métrica M5.3.1 (Total de registros carregados com sucesso na tabela de fatos)			
C24	Geração das medidas de qualidade para as métricas M6.1.1 e 13.1.1 (Nota para a atualidade)			
C25	Geração das medidas de qualidade para a métrica M9.1.1 (Rotinas de verificação que retornaram sucesso).			
C26	Geração das medidas de qualidade para a métrica M12.2.1 (Membros não categorizados nas hierarquias).			

ANEXO IV

ESPECIFICAÇÃO DOS COMPONENTES DA BIBLIOTECA

C01

Nome: C01 – Inclusão dos Objetivos Questionamentos e Métricas

Entradas: Objetivos, questionamentos e métricas

Saídas: Objetivos, questionamentos e m,étricas atualizadas no RMQ

Processamento:

- 1. Receber o objetivo
- 2. Incluir o objetivo no RMQ
- 3. Receber os questionamentos do objetivo
- 4. Para cada questionamento
 - 4.1. Incluir o questionamento no RMQ, associando-o ao objetivo
 - 4.2. Receber a métrica referente ao questionamento
 - 4.3. Incluir a métrica no RMQ associando-a ao questionamento

C02

Nome: Carga dos Objetos do DW no RMQ

Entradas: Objetos do DW não cadastrados no RMQ

Saídas: Objetos do DW atualizados no RMQ

Processamento:

- 1. Para cada objeto do DW ainda não cadastrado no RMQ (dimensão, fato, tupla ou célula)
 - 1.1. Obter o identificador do objeto no DW
 - 1.2. Incluir o objeto em OBJETO DO DW, no RMQ
 - 1.3. Se o objeto pertence ao nível 0 da arquitetura multicamadas, associá-lo ao objeto do qual ele é instância, através da associação "instância de"
 - 1.4. Se o objeto pertence ao nível 1 da arquitetura multicamadas, associá-lo ao tipo de objeto (ELEMENT) correspondente

C03

Nome: C03 – Cálculo do Objetivo

Entradas: Identificação do objetivo

Saídas: Lista de combinações (grau da qualidade, lista de objetos do DW)

Se a qualidade do objeto do DW não depende de outros objetos, a saída é apenas uma ocorrência de (grau da qualidade, lista de objetos), onde a lista de objetos é vazia.

Se a qualidade do objeto do DW depende de outros objetos, a lista conterá todas as combinações (grau da qualidade, lista dos objetos em relação aos quais o grau foi obtido).

- 1. Receber a identificação do objetivo
- 2. Se as medidas armazenadas no RMQ para o objetivo não dependem de outros

objetos do DW (ou seja, não estão relacionadas a nenhum objeto do DW através da associação "depende")

- 2a.1 Para cada métrica associada ao objetivo (via questionamento)
 - 2a.1.1 Obter a medida de qualidade armazenada no RMQ
 - 2a.2.1 Calcular valor uniformizado para medida, utilizando a equação (1) ou (2) da seção 4.7.3
- 2a.2 Obter o grau de qualidade para o objetivo, utilizando a equação (3) da seção 4.7.3
- 2a.3 Retornar a combinação (grau de qualidade, lista de objetos vazia)

Senão

- 2b.1 Agrupar as medidas que dependam de conjuntos de objetos do DW idênticos
- 2b.2 Para cada grupo formado
 - 2b.2.1 Para cada medida dependente do grupo de objetos 2b.2.1.1 Calcular valor uniformizado para medida, utilizando a equação (1) ou (2) da seção 4.7.3
 - 2b.2.2 Obter o grau de qualidade para o objetivo, utilizando a equação (3) da seção 4.7.3
 - 2b.2.3 Incluir a combinação (grau de qualidade, lista de objetos). (do DW) na lista de retorno
 - 2.b.3 Retornar a lista com todas as combinações (grau de qualidade, lista de). (objetos do DW)

C04

Nome: C04 - Carga da dimensão DIMENSAO

Entradas:

Saídas: Dimensão DIMENSAO atualizada

Processamento:

- 1. Consultar o repositório de metadados do DW para obter o nome das dimensões existentes.
- 2. Carregar as informações das dimensões na dimensão DIMENSAO.

C05

Nome: C05 - Carga da dimensão TEMPO

Entradas:

Saídas: Dimensão TEMPO atualizada

Processamento:

1. Incluir datas na dimensão TEMPO

C06

Nome: C06 - Carga da tabela de fatos do Cubo das Dimensões

Entradas:

Saídas: Tabela de fatos do Cubo das Dimensões atualizados

Processamento:

- Para cada cubo de fatos cuja referência esteja armazenada em OBJETO DO DW, no RMQ
 - 1.1.1. Para cada dimensão cuja referência esteja armazenada em OBJETO DO DW, no RMQ
 - 1.1.1.1.Enquanto houver CRITERIO DE QUALIDADE (Objetivo)
 - 1.1.1.1. Invocar o componente C03 para calcular o grau de qualidade para o objetivo, passando como parâmetro a identificação do mesmo. Como a qualidade de objetos do tipo "dimensão" não depende de outros objetos, o retorno de C03 será uma ocorrência (grau de qualidade, lista de objetos vazia).
 - 1.1.1.1.2. Armazenar o grau de qualidade na tabela de fatos do Cubo das Dimensões

C07

Nome: C07 - Carga da dimensão GRUPO DIMENSAO

Entradas:

Saídas: Dimensão GRUPO DIMENSAO atualizada

Processamento:

- 1. Criar um grupo para cada combinação possível de dimensões existentes na dimensão DIMENSAO
- 2. Armazenar os grupos na dimensão GRUPO DIMENSAO

C08

Nome: C08 - Carga da dimensão FATO

Entradas:

Saídas: Dimensão FATO atualizada

Processamento:

- 1. Consultar o repositório de metadados do DW para obter o nome dos fatos existentes.
- 2. Carregar as informações dos fatos na dimensão FATO.

C09

Nome: C09 - Carga da tabelas de fatos do Cubo dos Fatos

Entradas:

Saídas: Tabela de fatos do Cubo dos Fatos atualizada

- Para cada cubo de fatos cuja referência esteja armazenada em OBJETO DO DW, no RMQ
 - 1.1. Para cada fato cuja referência esteja armazenada em OBJETO DO DW, no RMQ
 - 1.1.1. Enquanto houver CRITERIO DE QUALIDADE (Objetivo)
 - 1.1.1.1.Invocar o componente C03 para calcular o grau de qualidade para o objetivo, passando como parâmetro, a identificação do mesmo. Como a qualidade de objetos do tipo "fato" depende das dimensões consideradas, o retorno do componente C03 será uma lista de pares (grau de qualidade, lista de dimensões em relação às quais a qualidade foi obtida).

- 1.1.1.2. Agrupar os pares da lista que possuam listas de dimensões idênticas
- 1.1.1.3.Para cada grupo formado
 - 1.1.1.3.1. Obter em GRUPO_DIMENSAO o número do grupo que corresponde à lista de dimensões
 - 1.1.1.3.2. Obter os graus de qualidade existentes nos pares do grupo e armazená-los na tabela de fatos do Cubo dos Fatos junto com o número de grupo obtido no passo anterior

C10

Nome: C10 - Carga da dimensão CUBO

Entradas:

Saídas: Dimensão CUBO atualizada

Processamento:

- 1. Consultar no repositório de metadados do DW para obter o nome dos cubos de dados.
- 2. Carregar as informações dos cubos na dimensão CUBO.

C11

Nome: C11 - Carga da tabela de auditoria das tuplas

Entradas:

Saídas: Tabela de auditoria das tuplas carregada

Processamento:

- 1. Para cada tabela de fatos pertencente a cubos de dados
 - 1.1. Para cada tupla da tabela de fatos, cuja referência esteja armazenada em OBJETO DO DW, no RMQ
 - 1.1.1. Enquanto houver CRITERIO DE QUALIDADE (Objetivo)
 - 1.1.1.1.Invocar o componente C03 para calcular o grau de qualidade para o objetivo, passando como parâmetro a identificação do mesmo. Como a qualidade de objetos do tipo "tupla" não depende de outros objetos, o retorno de C03 será uma ocorrência (grau de qualidade, lista de objetos vazia).
 - 1.1.1.2. Armazenar o grau de qualidade na tabela de auditoria das tuplas

C12

Nome: C12 - Carga da tabela de auditoria das células

Entradas:

Saídas: Tabela de auditoria das células atualizada

- 1. Para cada tabela de fatos pertencente a cubos de dados
 - 1.1. Para cada célula da tabela de fatos, cuja referência esteja armazenada em OBJETO DO DW, no RMQ
 - 1.1.1. Obter o fato do qual a célula é instância através do auto-relacionamento "instancia de" de OBJETO DO DW
 - 1.1.2. Enquanto houver CRITERIO DE QUALIDADE(Objetivo)
 - 1.1.2.1.Invocar o componente C01 para calcular o grau de qualidade para o objetivo, passando como parâmetro a identificação do mesmo. Como a qualidade de objetos do tipo "célula" não depende de outros

- objetos, o retorno de C01 será uma ocorrência (grau de qualidade, lista de objetos vazia).
- 1.1.2.2.Armazenar o grau de qualidade na tabela de auditoria das tuplas juntamente com o fato do qual a célula é instância

C13- Descrito no anexo V.

C14

Nome: C14

Descrição: Geração das medidas de qualidade para as métricas M1.1.1 e M10.1.1 (Registros que obedecem às regras de formação sintática).

Entradas:

- Registros (da tabela de fatos, para objetos do tipo "fato" ou da dimensão, para objetos do tipo "dimensão").
- Regras sintáticas armazenadas no repositório de metadados do DW.

Saídas: Valor de zero a cem, indicando o percentual de registros que obedecem às regras de formação sintática.

Processamento:

- 1. n = número total de regras sintáticas
- 2. Para cada regra de sintática i, referente ao objeto em questão, obtida do repositório de metadados do DW
 - 2.1. Para cada registro
 - 2.1.1. Se o atributo referente ao objeto obedece à regra sintática i 2.1.1.1. Incrementa contador "obedece à regra i"
 - 2.1.1.1. merementa contador obedece a regra r
 - 2.2. si = (contador "obedece à regra i" * 100) / quantidade total de registros
 - 2.3. wi = peso da regra sintática i
- 3. Retorna $((\sum_{i=1}^{n} si^*w_i) / n)$

C15

Nome: C15

Descrição: Geração das medidas de qualidade para as métricas M1.2.1 e M10.2.1 (Registros que pertencem ao domínio).

Entradas: Registros (da tabela de fatos, para objetos do tipo "fato" ou da dimensão, para objetos do tipo "dimensão").

Saídas: Valor de zero a cem, indicando o percentual de registros que pertencem ao domínio

- 1. Para cada registro
 - 1.1 Se o valor do atributo pertence ao domínio
 - 1.1.1. Incrementa contador "pertence ao domínio"
- 2. Retorna ((contador "pertence ao domínio" * 100) / quantidade total de registros)

C16

Nome: C16

Descrição: Geração das medidas de qualidade para as métricas M1.3.1 e M10.3.1 (Registros que pertencem ao tipo de dados).

Entradas: Registros (da tabela de fatos, para objetos do tipo "fato" ou da dimensão, para objetos do tipo "dimensão").

Saídas: Valor de zero a cem, indicando o percentual de registros que pertencem ao tipo de dado.

Processamento:

- 1. Para cada registro
 - 1.1. Se o valor do atributo pertence ao tipo de dado
 - 1.1.1. Incrementa contador "pertence ao tipo de dado"
- 2. Retorna ((contador "pertence ao tipo de dado"* 100) / quantidade total de registros)

C17

Nome: C17

Descrição: Geração das medidas de qualidade para as métricas M2.1.1 e M11.1.1 (Registros que obedecem às regras de negócio).

Entradas:

- Registros (da tabela de fatos, para objetos do tipo "fato" ou da dimensão, para objetos do tipo "dimensão").
- Regras de negócio armazenadas no repositório de metadados do DW

Saídas: Valor de zero a cem, indicando o percentual de registros que obedecem às regras de negócio.

Processamento:

- 1. n = número total de regras de negócio
- 2. Para cada regra de negócio i, referente ao objeto em questão, obtidas do repositório de metadados do DW
 - 2.1. Para cada registro
 - $2.1.1. \, \text{Se}$ o atributo referente ao objeto obedece à regra de negócio i
 - 2.1.1.1. Incrementa contador "obedece à regra i"
 - 2.2. ei = (contador "obedece à regra i" * 100) / quantidade total de registros
 - 2.3. wi = peso da regra de negócio i
- 3. Retorna $((\sum_{i=1}^{n} ei * w_i) / n)$

C18

Nome: C18

Descrição: Geração das medidas de qualidade para a métrica M2.2.1 (Registros contendo chaves estrangeiras cadastradas nas dimensões).

Entradas: - Registros da tabela de fatos

- Dimensões em relação às quais será medida a qualidade do fato

Saídas: Valor de zero a cem, indicando o percentual de registros que possuem chaves cadastradas nas dimensões.

Processamento:

1. n = número total de dimensões consideradas

- 2. Para cada dimensão i
 - 2.1. Para cada registro
 - 2.1.1. Se a chave estrangeira está cadastrada na dimensão
 - 2.1.1.1. Incrementa contador "chave cadastrada"
 - 2.2. qi = (contador "chave cadastrada" * 100) / quantidade total de registros
- 3. Retorna $((\sum_{i=1}^{n} qi) / n)$

C19

Nome: C19

Descrição: Geração das medidas de qualidade para a métrica M3.1.1

(Registros que estão em conformidade com fonte conhecida).

Entradas: Registros da tabela de fatos.

Saídas: Valor de zero a cem, indicando o percentual de registros que estão em conformidade com fonte conhecida.

Processamento:

- 1. Para cada registro
 - 1.1. Compara campo com fonte conhecida
 - 1.2. Se o campo é igual ao valor da fonte conhecida
 - 1.2.1. Incrementa contador "igual à fonte"
- 2. Retorna ((contador "igual à fonte" * 100) / quantidade total de registros)

C20

Nome: C20

Descrição: Geração das medidas de qualidade para a métrica M4.1.1

(Proximidade do valor armazenado em relação ao valor real).

Entradas: - Tipo de aproximação

- Regra para geração do valor aproximado

Saídas: Valor de zero a cem, indicando proximidade percentual do valor armazenado em relação ao real (quanto maior o percentual, mais próximo o valor aproximado está do valor real).

Processamento:

1. Retorna um valor referente ao percentual de proximidade, em função da regra de geração e do tipo de aproximação.

C21

Nome: C21

Descrição: Geração das medidas de qualidade para a métrica M5.1.1

(Registros onde o campo está preenchido ou está identificado como não aplicável).

Entradas: Registros da tabela de fatos.

Saídas: Valor de zero a cem, indicando o percentual de registros preenchidos.

Processamento:

- 1. Para cada registro
 - 2. Se o campo está preenchido ou está identificado como não aplicável
 - 2.1. Incrementa contador "preenchido"

2. Retorna ((contador "preenchido" * 100) / quantidade total de registros)

C22

Nome: C22

Descrição: Geração das medidas de qualidade para a métrica M5.2.1 e M12.1.1 (Membros existentes que estão cadastrados nas dimensões).

Entradas: - Dimensões consideradas

- Membros existentes em cada dimensão considerada

Saídas: Valor de zero a cem, indicando o percentual membros cadastrados nas dimensões.

Processamento:

- 1. n = número total de dimensões consideradas
- 2. Para cada dimensão i
 - 2.1. Para cada membro
 - 2.1.1. Se membro está cadastrado na dimensão
 - 2.1.1.1. Incrementa contador "membro cadastrado"
 - 2.2. qi = (contador "membro cadastrado" * 100) / quantidade total de membros
- 3. Retorna $((\sum_{i=1}^{n} qi) / n)$

C23

Nome: C23

Descrição: Geração das medidas de qualidade para a métrica M5.3.1

(Total de registros carregados com sucesso na tabela de fatos)

Entradas: - Data da carga

Saídas: Valor de zero a cem, indicando o percentual de registros carregados

Processamento:

- 1. Consulta o *log* da carga para a data especificada e obtém a quantidade total de registros e a quantidade de registros carregados
- 2. Retorna ((quantidade de registros carregados"* 100) / quantidade total de registros)

C24

Nome: C24

Descrição: Geração das medidas de qualidade para as métricas M6.1.1 e 13.1.1 (Nota para a atualidade)

Entradas: - Data da última carga

- Periodicidade de carga

Saídas: Zero ou cem, em função da atualidade.

Processamento:

- 1. Se data da última carga está dentro da periodicidade de carga
 - 1.1. Retorna (100)

Senão

1.2. Retorna (0)

C25

Nome: C25

Descrição: Geração das medidas de qualidade para a métrica M9.1.1

(Rotinas de verificação que retornaram sucesso).

Entradas: Rotinas de verificação.

Saídas: Valor de zero a cem, indicando o percentual de rotinas de verificação que retornaram sucesso.

Processamento:

- 1. Para cada rotina de verificação
 - 1.1. Executa rotina
 - 1.2. Se rotina retornou sucesso
 - 1.2.1. Incrementa contador "sucessos"
- 2. Retorna ((contador "sucessos" * 100) / quantidade total de rotinas)

C26

Nome: C26

Descrição: Geração das medidas de qualidade para a métrica M12.2.1

(Membros não categorizados nas hierarquias).

Entradas: - Dimensão

- Membros existentes na dimensão

Saídas: Valor de zero a cem, indicando o percentual de membros não categorizados nas hierarquias

Processamento:

- 1. Para cada membro
 - 1.1. Se membro não está categorizado nas hierarquias
 - 1.1.1. Incrementa contador "membro não categorizado"
- 2. Retorna((contador "membro não categorizado" * 100) / quantidade total de membros)

ANEXO V

ESPECIFICAÇÃO DOS MÉTODOS DISCOVER E EXECUTE DO XMLA-Q

MÉTODO DISCOVER DO XMLA-Q

Componente: C13 – Serviço de Qualidade

Nome do método: Discover

Sintaxe:

Execute (

[in] *RequestType* As EnumString,

[in] Restrictions As Restrictions,

[in] *Properties* As Properties,

[out] Result As Rowset

Entradas:

ResquestType: determina o tipo de informação que será retornada. O parâmetro de mesmo nome do XMLA original (MICROSOFT, HYPERION, 2002) foi estendido para aceitar valores correspondentes a solicitações de metadados dos cubos de qualidade. São eles:

QSCHEMA_CUBES Enumeração dos Cubos de Qualidade

QSCHEMA_DIMENSIONS Enumeração das Dimensões de todos os Cubos de

Qualidade

QSCHEMA_MEASURES Enumeração das Medidas em todos os Cubos de

Qualidade

Consequentemente, o tipo de objeto *Rowset* do XMLA (MICROSOFT, HYPERION, 2002) também foi estendido para contemplar os novos tipos de *RequestType*. Foram acrescentados ao *Rowset* do XMLA, os seguintes tipos:

QSCHEMA_CUBES Rowset: composto pelo elemento CUBE_NAME que representa o nome do cubo

QSCHEMA_DIMENSIONS Rowset: composto pelos elementos CUBE_NAME e

DIMENSION_NAME, que representam o nome do cubo e da dimensão, respectivamente.

QSCHEMA_MEASURES Rowset: composto pelos elementos CUBE_NAME e

MEASURE_NAME, que representam o nome do

cubo e da medida, respectivamente.

Restrictions: possibilita que sejam estabelecidos filtros sobre os dados que devem ser retornados em *Result*. No XMLA-Q o tipo Restrictions foi estendido para suportar os novos tipos de *Rowset*.

Properties: consiste de uma coleção de *Properties*. No método *Discover* do XMLA-Q, esse parâmetro assume os mesmos valores que no XMLA (MICROSOFT, HYPERION, 2002).

Saída:

Result: Contém o resultado retornado como um objeto Rowset definido no XMLA-Q.

Processamento:

- 1. Receber a solicitação.
- 2. Se RequestType for igual a

QSCHEMA_CUBES ou QSCHEMA_DIMENSIONS ou QSCHEMA_MEASURES

Então

2.a.1 Retornar para o cliente, respectivamente:

Metadados dos Cubos de Qualidade ou

Metadados das Dimensões de todos os Cubos de Qualidade ou

Metadados das Medidas de todos os Cubos de Qualidade

Senão

- 2.b.1 Repassar a chamada para o método Discovery do Serviço de Consulta
- 2.b.2 Redirecionar a resposta recebida para o cliente

Exemplo:

1. Solicitação do nome dos Cubos de Qualidade:

Chamada do método Discover

```
<SOAP-ENV: Envelope
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV:Body>
 <Discover xmlns="urn:schemas-microsoft-com:xml-analysis"</pre>
 SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
   <RequestType>QSCHEMA_CUBES</RequestType>
   <Properties>
    <PropertyList>
     <DataSourceInfo>
     Provider=MSOLAP;Data Source=local;
     </DataSourceInfo>
     <Format>
     Tabular
     </Format>
    </PropertyList>
```

```
</Properties>
</Discover> </SOAP-ENV:Body> </SOAP-ENV:Envelope>
```

Resposta do método Discover:

```
<?xml version="1.0"?>
<SOAP-ENV:Envelope
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV: Body>
 <DiscoverResponse xmlns="urn:schemas-microsoft-com:xml-analysis">
  <return>
       <root xmlns="urn:schemas-microsoft-com:xml-analysis:rowset">
            <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
            <!-Definição do esquema XML do resultado estará aqui -->
            </xsd:schema>
            <row>
                <CUBE_NAME>Cubo_Dimensoes
            </row>
                <CUBE_NAME>Cubo_Fatos
            </row>
       </root>
  </return>
 </DiscoverResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

2. Solicitação do nome das Dimensões dos Cubos de Qualidade:

Chamada do método Discover:

```
<SOAP-ENV: Envelope
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV: Body>
 <Discover xmlns="urn:schemas-microsoft-com:xml-analysis"</pre>
 SOAP-ENV: encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
  <RequestType>QSCHEMA_DIMENSIONS</RequestType>
   <Properties>
   <PropertyList>
     <DataSourceInfo>
     Provider=MSOLAP; Data Source=local;
    </DataSourceInfo>
    <Format>
     Tabular
    </Format>
    </PropertyList>
   </Properties>
</Discover> </SOAP-ENV:Body> </SOAP-ENV:Envelope>
```

Resposta do método Discover:

```
<?xml version="1.0"?>
<SOAP-ENV:Envelope
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV:Body>
```

```
<DiscoverResponse xmlns="urn:schemas-microsoft-com:xml-analysis">
  <return>
       <root xmlns="urn:schemas-microsoft-com:xml-analysis:rowset">
            <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
            <!-Definição do esquema XML do resultado estará aqui -->
            </xsd:schema>
                <CUBE_NAME>Cubo_Dimensoes</CUBE_NAME>
                <DIMENSION_NAME>Dimensao<DIMENSION_NAME>
            </row>
            <row>
                <CUBE_NAME>Cubo_Dimensoes
                <DIMENSION_NAME>Cubo<DIMENSION_NAME>
            </row>
            <row>
                <CUBE_NAME>Cubo_Dimensoes
                <DIMENSION_NAME>Tempo<DIMENSION_NAME>
            < r0w>
                <CUBE_NAME>Cubo_Fatos</CUBE_NAME>
                <DIMENSION_NAME>Grupo_Dimensao<DIMENSION_NAME>
            </row>
                <CUBE_NAME>Cubo_Fatos
                <DIMENSION_NAME>Fato<DIMENSION_NAME>
            </row>
            <row>
                <CUBE_NAME>Cubo_Fatos</CUBE_NAME>
                <DIMENSION_NAME>Cubo<DIMENSION_NAME>
            </row>
            <row>
                <CUBE_NAME>Cubo_Fatos
                <DIMENSION_NAME>Tempo<DIMENSION_NAME>
            </row>
       </root>
  </return>
 </DiscoverResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

3. Solicitação do nome das Medidas dos Cubos de Qualidade:

Chamada do método Discover:

```
<SOAP-ENV: Envelope
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV:Body>
<Discover xmlns="urn:schemas-microsoft-com:xml-analysis"</pre>
 SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
  <RequestType>QSCHEMA_MEASURES
  <Properties>
   <PropertyList>
    <DataSourceInfo>
     Provider=MSOLAP; Data Source=local;
    </DataSourceInfo>
    <Format>
     Tabular
    </Format>
   </PropertyList>
  </Properties>
</Discover> </SOAP-ENV:Body> </SOAP-ENV:Envelope>
```

Resposta do método Discover:

```
<?xml version="1.0"?>
<SOAP-ENV: Envelope
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV:Body>
 <DiscoverResponse xmlns="urn:schemas-microsoft-com:xml-analysis">
  <return>
       <root xmlns="urn:schemas-microsoft-com:xml-analysis:rowset">
            <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
            <!-Definição do esquema XML do resultado estará aqui -->
            </xsd:schema>
                <CUBE_NAME> Cubo_Dimensoes/CUBE_NAME>
                <MEASURE_NAME>Grau_Qualidade1<MEASURE_NAME>
            </row>
            <row>
                <CUBE_NAME> Cubo_Dimensoes/CUBE_NAME>
                <MEASURE_NAME>Grau_Qualidade2<MEASURE_NAME>
            </row>
                <CUBE_NAME> Cubo_Fatos
                <MEASURE_NAME>Grau_Qualidade1<MEASURE_NAME>
            <row>
                <CUBE_NAME> Cubo_Fatos
                <MEASURE_NAME>Grau_Qualidade2<MEASURE_NAME>
            </row>
       </root>
  </return>
 </DiscoverResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

O MÉTODO EXECUTE DO XMLA-Q

```
Nome do método: Execute

Sintaxe:

Execute (

[in] Command As Command,

[in] Properties As Properties,

[out] Result As ResultSet
```

Entradas:

Command: executa uma consulta em MDX. Consiste de uma sentença na sintaxe MDX, encapsulada pelo elemento <Statement> do mdXML.

Properties: controla aspectos do retorno do método. Análogo ao parâmetro de mesmo

nome, do XMLA (MICROSOFT, HYPERION, 2002), porém contendo as seguintes alterações para o método *Execute*:

- a opção "Content" foi alterada, para estar adequada ao contexto de qualidade. O enumerador que especifica os valores possíveis para esta opção foi redefinido, podendo assumir os seguintes valores:
 - Q_DIM: significa que devem ser retornados dados de qualidade sobre dimensões, contidos no Cubo das Dimensões
 - Q_DIM_A: além dos dados de qualidade sobre dimensões, retorna também as anotações sobre as mesmas, contidas no Cubo de Anotações das Dimensões.
 - Q_FACTS: significa que devem ser retornados dados de qualidade sobre fatos, contidos no Cubo dos Fatos
 - Q_FACTS_A: além dos dados de qualidade sobre fatos, retorna também as anotações sobre os mesmos, contidas no Cubo de Anotações dos Fatos.
 - Q_TUPLES: significa que devem ser retornados dados de qualidade sobre tuplas, contidos nos Cubos das Tuplas
 - Q_TUPLES_A: além dos dados de qualidade sobre tuplas, retorna também as anotações sobre as mesmas, contidas no Cubo de Anotações de Tuplas.
 - Q_CELLS: significa que devem ser retornados dados de qualidade sobre células, contidos nos Cubos das Células
 - Q_CELLS_A: além dos dados de qualidade sobre células, retorna também as anotações sobre as mesmas, contidas no Cubo de Anotações das Células.
- criação da opção "QualityRestrictions", do tipo Restrictions do XMLA-Q.
 Possibilita que seja feita uma seleção das medidas dos Cubos de Qualidade que
 devem ser retornadas. As restrições devem ser especificadas no formato
 QSCHEMA_MEASURES Rowset. Se não forem especificadas restrições, serão
 retornadas todas as medidas dos Cubos de Qualidade envolvidos.

A opção "QualityRestrictions" só deve ser utilizada caso o conteúdo de "Content" seja igual a Q_DIM, Q_DIM_A, Q_FACTS, Q_FACTS_A, Q_TUPLES_A, Q_CELLS ou Q_CELLS_A.

Saída:

Result: possui o mesmo formato que o parâmetro de mesmo nome, no XMLA (MICROSOFT, HYPERION, 2002). Dependendo da solicitação definida em *Properties*, retorna dados propriamente ditos ou informações extraídas dos cubos de qualidade (referentes à qualidade dos dados que seriam retornados como resultado da

consulta especificada em Command).

Processamento:

- 1. Receber de uma consulta em mdXML, enviada pelo cliente;
- 2. Se Properties possui "Content" igual a

```
Q_DIM ou
Q_DIM_A ou
Q_FACTS ou
Q_FACTS_A ou
Q_TUPLES ou
Q_TUPLES_A ou
Q_CELLS ou
Q_CELLS A
```

Então

- 2a.1 Extrair os objetos envolvidos na consulta
- 2a.2 Montar uma consulta (em mdXML) aos Cubos de Qualidade, respeitando as restrições definidas na opção "QualityRestrictions" de *Properties*, selecionando apenas as informações de qualidade dos objetos identificados em 2a.1
- 2a.3 Enviar a consulta para o Serviço de Consulta, usando XMLA
- 2a.4 Receber o resultado e redirecioná-lo para o cliente

Senão

- 2b.1 Enviar a consulta para o Serviço de Consulta, usando XMLA
- 2b.2 Receber o resultado e redirecioná-lo para o cliente

Exemplo:

Neste exemplo, o cliente faz uma chamada ao método *Execute*, solicitando o Grau Qualidade1 das dimensões envolvidas na consulta:

```
SELECT [MEASURES].MEMBERS ON COLUMNS [LOJA].MEMBERS ON ROWS FROM VENDAS
```

O Serviço de Qualidade monta a consulta de qualidade abaixo e a envia para o Servidor de Consultas:

```
SELECT MEASURES.[GRAU_QUALIDADE1] ON COLUMNS [DIMENSAO].[LOJA] ON ROWS FROM CUBO_DIMENSOES WHERE (TTEMPO.[10102003], CUBO.[VENDAS])
```

O Servidor de consultas envia a resposta, que é redirecionada para o cliente pelo Serviço de Qualidade, contendo o Grau_Qualidade1 da dimensão "Loja", na data

"10/10/2003" (obtido do Cubo das Dimensões). A seguir são apresentados os comandos envolvidos nessa consulta:

Chamada do método Execute:

```
<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV: encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"> <SOAP-
ENV: Body>
<Execute xmlns="urn:schemas-microsoft-com:xml-analysis"</pre>
  SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
   <Command>
    <Statement>
     select [Measures].members on Columns
     [Loja].Members on Rows
     from Vendas
    </Statement>
   <Command>
 <Properties>
    <PropertyList>
     <DataSourceInfo>
        Provider=Essbase; Data Source=local;
     </DataSourceInfo>
     <Content>Q DIM</Content>
     <QualityRestrictions>
             <RestrictionList>
              <CUBE NAME>Cubo Dimensoes</CUBE NAME>
              <MEASURE_NAME>Grau_Qualidade1<MEASURE_NAME>
             </RestrictionList>
     </QualityRestrictions>
     <Format>Multidimensional/Format>
     <AxisFormat>ClusterFormat</AxisFormat>
    </PropertyList>
</Properties>
</Execute>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
<?xml version="1.0"?>
Resposta do método Execute:
<SOAP-ENV: Envelope xmlns: SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV: encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"> <SOAP-
ENV: Body>
<m:ExecuteResponse xmlns:m="urn:schemas-microsoft-com:xml-analysis">
<m:return SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
       <root xmlns="urn:schemas-microsoft-com:xml-analysis:mddataset">
          <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"</pre>
             xmlns:xars="urn:schemas-microsoft-com:xars">
             ...<!-Esquema XML dos dados estará aqui. -- >
          </xsd:schema>
          ... <!-Dados no formato MDDataSet (MICROSOFT, HYPERION, 2002)
                 estarão aqui. -- >
       </root>
</m:return>
</m:ExecuteResponse>
 </SOAP-ENV:Body>
 </SOAP-ENV:Envelope>
```

ANEXO VI

COMANDOS XMLA DA CONSULTA DO EXEMPLO DE APLICAÇÃO

Os resultados das consultas (*Result*) são apresentados no formato *MDDataSet* do XMLA. A especificação do XMLA (MICROSOFT, HYPERION, 2002) contém a descrição detalhada deste formato.

1. Solicitação de informações sobre a qualidade das dimensões envolvidas

```
Chamada do método Execute do XMLA-Q:
<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"> <SOAP-
ENV:Body>
< Execute xmlns="urn:schemas-microsoft-com:xml-analysis"
 SOAP-ENV: encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
   <Command>
      <Statement>
          SELECT {[MEASURES].[QTD_CRIMES_ENTORPECENTES],
                  [MEASURES].[QTD_PACIENTES_ENTORPECENTES],
                  [MEASURES].[QTD_ALUNOS_ENTORPECENTES] } ON COLUMNS,
                  [MUNICIPIOS].MEMBERS ON ROWS
           FROM
                  SEGURANCA_PUBLICA
           WHERE [TEMPO].[2003]
      </Statement>
   </Command>
   <Properties>
      <PropertyList>
         <DataSourceInfo>
            Provider=Essbase;Data Source=local;
         </DataSourceInfo>
         <Content>Q_DIM_A</Content>
         <Format>Multidimensional</Format>
         <AxisFormat>TupleFormat</AxisFormat>
      </PropertyList>
   </Properties>
</Execute>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
<?xml version="1.0"?>
Resposta do método Execute do XMLA-Q (formato MDDataSet):
<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV: encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"> <SOAP-
ENV: Body>
<m:ExecuteResponse xmlns:m="urn:schemas-microsoft-com:xml-analysis">
<m:return SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
       <root xmlns="urn:schemas-microsoft-com:xml-analysis:mddataset">
          <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"</pre>
             xmlns:xars="urn:schemas-microsoft-com:xars">
             ...<!-Esquema XML dos dados estará aqui. -- >
          </xsd:schema>
<!- Inicio dos Dados no formato MDDataSet-- >
```

```
<OlapInfo>
      <CubeInfo>
             <Cube>
                    <CubeName>Qualidade_Dimensoes</CubeName>
             </Cube>
      </CubeInfo>
      <AxesInfo>
             <AxisInfo name="Axis0">
                    <hi>HierarchyInfo name="Measures"></hi>
                           <UName name="[Measures].[MEMBER_UNIQUE_NAME]"></UName>
                           <Caption name="[Measures].[MEMBER_CAPTION]"></Caption>
                           <LName name="[Measures].[LEVEL_UNIQUE_NAME]"></LName>
                           <LNum name="[Measures].[LEVEL_NUMBER]"></LNum>
                           <DisplayInfo
                           name="[Measures].[DISPLAY_INFO]"></DisplayInfo>
                    </HierarchyInfo>
             </AxisInfo>
             <AxisInfo name="Axis1">
                    <HierarchyInfo name="Dimensoes">
                           <UName name="[Dimensoes].[MEMBER_UNIQUE_NAME]"></UName>
                           <Caption name="[Dimensoes].[MEMBER_CAPTION]"></Caption>
                           <LName name="[Dimensoes].[LEVEL_UNIQUE_NAME]"></LName>
                           <LNum name="[Dimensoes].[LEVEL_NUMBER]"></LNum>
                           <DisplayInfo
                           name="[Dimensoes].[DISPLAY_INFO]"></DisplayInfo>
                    </HierarchyInfo>
             </AxisInfo>
             <AxisInfo name="SlicerAxis">
                    <HierarchyInfo name="Tempo">
                           <UName name="[Tempo].[MEMBER_UNIQUE_NAME]"></UName>
                           <Caption name="[Tempo].[MEMBER_CAPTION]"></Caption>
                           <LName name="[Tempo].[LEVEL_UNIQUE_NAME]"></LName>
                           <LNum name="[Tempo].[LEVEL_NUMBER]"></LNum>
                           <DisplayInfo
                           name="[Tempo].[DISPLAY_INFO]"></DisplayInfo>
                    </HierarchyInfo>
             </AxisInfo>
      </AxesInfo>
      <CellInfo>
             <Value name="VALUE"></Value>
      </CellInfo>
</OlapInfo>
<Axes>
      <Axis name="Axis0">
             <Tuples>
                    <Tuple>
                           <Member Hierarchy="Measures">
                                 <UName>[Measures].[acuracia_sintatica]</UName>
                                 <Caption>Acurácia Sintatica</Caption>
                                 <LName>[Measures].[MeasuresLevel]</LName>
                                 <LNum>0</LNum>
                           </Member>
                    </Tuple>
                    <Tuple>
                           <Member Hierarchy="Measures">
                                 <UName>[Measures].[acuracia_semantica]</UName>
                                 <Caption>Acurácia Semantica</Caption>
                                 <LName>[Measures].[MeasuresLevel]</LName>
                                 <LNum>0</LNum>
                           </Member>
                    </Tuple>
                    <Tuple>
                           <Member Hierarchy="Measures">
```

```
<UName>[Measures].[completeza]</UName>
                                 <Caption>Completeza</Caption>
                                 <LName>[Measures].[MeasuresLevel]</LName>
                                 <LNum>0</LNum>
                          </Member>
                    </Tuple>
                    <Tuple>
                          <Member Hierarchy="Measures">
                                 <UName>[Measures].[atualidade]</UName>
                                 <Caption>Atualidade</Caption>
                                 <LName>[Measures].[MeasuresLevel]</LName>
                                 <LNum>0</LNum>
                          </Member>
                    </Tuple>
                    <Tuple>
                          <Member Hierarchy="Measures">
                                 <UName>[Measures].[confiabilidade]</UName>
                                 <Caption>Confiabilidade</Caption>
                                 <LName>[Measures].[MeasuresLevel]</LName>
                                 <LNum>0</LNum>
                          </Member>
                    </Tuple>
                    <Tuple>
                          <Member Hierarchy="Measures">
                                 <UName>[Measures].[anotacao]</UName>
                                 <Caption>Anotacao</Caption>
                                 <LName>[Measures].[MeasuresLevel]
                                 <LNum>0</LNum>
                          </Member>
                    </Tuple>
             </Tuples>
      </Axis>
      <Axis name="Axis1">
             <Tuples>
                    <Tuple>
                          <Member Hierarchy="Dimensoes">
                                 <UName>[Dimensoes].[Municipio]</UName>
                                 <Caption>Municipio</Caption>
                                 <LName>[Dimensoes].[Municipio]</LName>
                                 <LNum>1</LNum>
                          </Member>
                    </Tuple>
             </Tuples>
      </Axis>
      <Axis name="SlicerAxis">
             <Tuples>
                    <Tuple>
                          <Member Hierarchy="Tempo">
                                 <UName>[Tempo].[2003]</UName>
                                 <Caption>2003</Caption>
                                 <LName>[Tempo].[Ano]</LName>
                                 <LNum>0</LNum>
                          </Member>
                    </Tuple>
             </Tuples>
      </Axis>
</Axes>
<CellData>
      <Cell CellOrdinal="0">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="1">
             <Value xsi:type="xsd:double">100</Value>
```

```
</Cell>
      <Cell CellOrdinal="2">
             <Value xsi:type="xsd:double">100</Value>
      <Cell CellOrdinal="3">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="4">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="5">
             <Value xsi:type="xsd:string">
                    <root>
                           <xsd:schema xmlns:xsd=http://www.w3.org/2001/XMLSChema>
                          <!- O esquema XML da anotacao pode ser incluído aqui-->
                          </xsd:schema>
                          <GrausDeQualidade>
                                <Reputacao>Fonte: PRODERJ</Reputação>
                          </GrausDeQualidade>
                    </root>
             </Value>
      </Cell>
</CellData>
<!- Fim dos Dados no formato MDDataSet-- >
       </root>
</m:return>
</m:ExecuteResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Observações:

Durante a execução do método Execute, o Serviço de Qualidade traduz a consulta original para as seguintes consultas aos Cubos de Qualidade:

Comando MDX para seleção dos graus de qualidade das dimensões envolvidas:

```
SELECT [MEASURES].MEMBERS ON COLUMNS
[DIMENSAO].[MUNICIPIOS] ON ROWS
FROM CUBO_DIMENSOES
WHERE (TEMPO.[10112003], CUBO.[SEGURANCA_PUBLICA])
```

Comando MDX para seleção das anotações sobre as dimensões envolvidas:

```
SELECT [MEASURES].MEMBERS ON COLUMNS
[DIMENSAO].[MUNICIPIOS] ON ROWS
FROM ANOTACAO_DIMENSOES
WHERE (TEMPO.[10112003], CUBO.[SEGURANCA_PUBLICA])
```

2. Solicitação de informações sobre a qualidade dos fatos envolvidos

```
Chamada do método Execute do XMLA-Q:
<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/" SOAP-
ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"> <SOAP-ENV:Body>
<Execute xmlns="urn:schemas-microsoft-com:xml-analysis"</pre>
  SOAP-ENV: encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
   <Command>
      <Statement>
          SELECT {[MEASURES].[QTD_CRIMES_ENTORPECENTES],
                   [MEASURES].[QTD_PACIENTES_ENTORPECENTES],
                  [MEASURES].[QTD_ALUNOS_ENTORPECENTES] } ON COLUMNS,
                  [MUNICIPIOS].MEMBERS ON ROWS
           FROM
                  SEGURANCA PUBLICA
           WHERE [TEMPO].[2003]
      </Statement>
   <Command>
   <Properties>
      <PropertyList>
        <DataSourceInfo>
           Provider=Essbase; Data Source=local;
        </DataSourceInfo>
        <Content>Q_FACTS_A</Content>
        <Format>Multidimensional</Format>
        <AxisFormat>TupleFormat</AxisFormat>
      </PropertyList>
   </Properties>
</Execute>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
<?xml version="1.0"?>
Resposta do método Execute do XMLA-Q (formato MDDataSet)::
<SOAP-ENV:Envelope xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/" SOAP-
ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"> <SOAP-ENV:Body>
<m:ExecuteResponse xmlns:m="urn:schemas-microsoft-com:xml-analysis">
<m:return SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
       <root xmlns="urn:schemas-microsoft-com:xml-analysis:mddataset">
          <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"</pre>
             xmlns:xars="urn:schemas-microsoft-com:xars">
             ...<!-Esquema XML dos dados estará aqui. -- >
          </xsd:schema>
<!- Inicio dos dados no formato MDDataSet -- >
<OlapInfo>
       <CubeInfo>
             <Cube>
                    <CubeName>Qualidade_Fatos</CubeName>
             </Cube>
       </CubeInfo>
       <AxesInfo>
             <AxisInfo name="Axis0">
                    <HierarchyInfo name="Measures">
                           <UName name="[Measures].[MEMBER_UNIQUE_NAME]"></UName>
                           <Caption name="[Measures].[MEMBER_CAPTION]"></Caption>
                           <LName name="[Measures].[LEVEL_UNIQUE_NAME]"></LName>
                           <LNum name="[Measures].[LEVEL_NUMBER]"></LNum>
                           <DisplayInfo
```

```
name="[Measures].[DISPLAY_INFO]"></DisplayInfo>
                    </HierarchyInfo>
             </AxisInfo>
             <AxisInfo name="Axis1">
                    <HierarchyInfo name="Fatos">
                           <UName name="[Fatos].[MEMBER_UNIQUE_NAME]"></UName>
                           <Caption name="[Fatos].[MEMBER_CAPTION]"></Caption>
                           <LName name="[Fatos].[LEVEL_UNIQUE_NAME]"></LName>
                           <LNum name="[Fatos].[LEVEL_NUMBER]"></LNum>
                           <DisplayInfo
                          name="[Fatos].[DISPLAY_INFO]"></DisplayInfo>
                    </HierarchyInfo>
             </AxisInfo>
             <AxisInfo name="SlicerAxis">
                    <HierarchyInfo name="Tempo">
                           <UName name="[Tempo].[MEMBER_UNIQUE_NAME]"></UName>
                           <Caption name="[Tempo].[MEMBER_CAPTION]"></Caption>
                           <LName name="[Tempo].[LEVEL_UNIQUE_NAME]"></LName>
                           <LNum name="[Tempo].[LEVEL_NUMBER]"></LNum>
                           <DisplayInfo
                          name="[Tempo].[DISPLAY_INFO]"></DisplayInfo>
                    </HierarchyInfo>
             </AxisInfo>
      </AxesInfo>
      <CellInfo>
             <Value name="VALUE"></Value>
      </CellInfo>
</OlapInfo>
<Axes>
      <Axis name="Axis0">
             <Tuples>
                    <Tuple>
                           <Member Hierarchy="Measures">
                                 <UName>[Measures].[acuracia_sintatica]</UName>
                                 <Caption>Acurácia Sintatica</Caption>
                                 <LName>[Measures].[MeasuresLevel]</LName>
                                  <LNum>0</LNum>
                           </Member>
                    </Tuple>
                    <Tuple>
                           <Member Hierarchy="Measures">
                                 <UName>[Measures].[acuracia_semantica]</UName>
                                 <Caption>Acurácia Semantica</Caption>
                                 <LName>[Measures].[MeasuresLevel]</LName>
                                 <T_1N_{11}m>0</T_1N_{11}m>
                           </Member>
                    </Tuple>
                    <Tuple>
                           <Member Hierarchy="Measures">
                                 <UName>[Measures].[acuracia_conteudo]</UName>
                                 <Caption>Acurácia de Conteudo</Caption>
                                 <LName>[Measures].[MeasuresLevel]</LName>
                                 <LNum>0</LNum>
                           </Member>
                    </Tuple>
                    <Tuple>
                           <Member Hierarchy="Measures">
                                 <UName>[Measures].[precisao]</UName>
                                 <Caption>Precisao</Caption>
                                  <LName>[Measures].[MeasuresLevel]</LName>
                                 <LNum>0</LNum>
                           </Member>
                    </Tuple>
```

```
<Tuple>
                    <Member Hierarchy="Measures">
                          <UName>[Measures].[completeza]</UName>
                          <Caption>Completeza</Caption>
                          <LName>[Measures].[MeasuresLevel]</LName>
                          <LNum>0</LNum>
                    </Member>
             </Tuple>
             <Tuple>
                    <Member Hierarchy="Measures">
                          <UName>[Measures].[atualidade]</UName>
                          <Caption>Atualidade</Caption>
                           <LName>[Measures].[MeasuresLevel]</LName>
                          <LNum>0</LNum>
                    </Member>
             </Tuple>
             <Tuple>
                    <Member Hierarchy="Measures">
                          <UName>[Measures].[confiabilidade]</UName>
                          <Caption>Confiabilidade</Caption>
                          <LName>[Measures].[MeasuresLevel]</LName>
                          <LNum>0</LNum>
                    </Member>
             </Tuple>
             <Tuple>
                    <Member Hierarchy="Measures">
                          <UName>[Measures].[anotacao]</UName>
                          <Caption>Anotacao</Caption>
                           <LName>[Measures].[MeasuresLevel]
                          <LNum>0</LNum>
                    </Member>
             </Tuple>
      </Tuples>
</Axis>
<Axis name="Axis1">
      <Tuples>
             <Tuple>
                    <Member Hierarchy="Fatos">
                          <UName>[Fatos].[QTD_CRIMES_ENTORPECENTES]</UName>
                           <Caption>Qtd. Crimes Entorpecentes</Caption>
                          <LName>[Fatos].[QTD_CRIMES_ENTORPECENTES]</LName>
                          <LNum>1</LNum>
                    </Member>
             </Tuple>
             <Tuple>
                    <Member Hierarchy="Fatos">
                          <UName>[Fatos].[QTD_PACIENTES_ENTORPECENTES]</UName>
                          <Caption>Qtd. Pacientes Entorpecentes</Caption>
                          <LName>[Fatos].[QTD_PACIENTES_ENTORPECENTES]</LName>
                           <LNum>1</LNum>
                    </Member>
             </Tuple>
             <Tuple>
                    <Member Hierarchy="Fatos">
                          <UName>[Fatos].[QTD_ALUNOS_ENTORPECENTES]</UName>
                           <Caption>Qtd. Alunos Entorpecentes</Caption>
                          <LName>[Fatos].[QTD_ALUNOS_ENTORPECENTES]</LName>
                          <LNum>1</LNum>
                    </Member>
             </Tuple>
      </Tuples>
</Axis>
<Axis name="SlicerAxis">
      <Tuples>
```

```
<Tuple>
                          <Member Hierarchy="Tempo">
                                 <UName>[Tempo].[2003]</UName>
                                 <Caption>2003</Caption>
                                 <LName>[Tempo].[Ano]</LName>
                                 <LNum>0</LNum>
                          </Member>
                    </Tuple>
             </Tuples>
      </Axis>
<CellData>
      <Cell CellOrdinal="0">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="1">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="2">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="3">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="4">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="5">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="6">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="7">
             <Value xsi:type="xsd:string">
                   <root>
                          <xsd:schema xmlns:xsd=http://www.w3.org/2001/XMLSChema>
                          <!- O esquema XML da anotacao pode ser incluído aqui-->
                          </xsd:schema>
                          <GrausDeQualidade>
                               <Reputacao>Fonte: Delegacias Civis/Reputação>
                          </GrausDeQualidade>
                   </root>
             </Value>
      </Cell>
      <Cell CellOrdinal="8">
             <Value xsi:type="xsd:double">100</Value>
      <Cell CellOrdinal="9">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="10">
             <Value xsi:type="xsd:double">70</Value>
      </Cell>
      <Cell CellOrdinal="11">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="12">
             <Value xsi:type="xsd:double">80</Value>
      </Cell>
      <Cell CellOrdinal="13">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="14">
             <Value xsi:type="xsd:double">100</Value>
```

```
</Cell>
             <Value xsi:type="xsd:string">
                    <root>
                           <xsd:schema xmlns:xsd=http://www.w3.org/2001/XMLSChema>
                          <!- O esquema XML da anotacao pode ser incluído aqui-->
                          </xsd:schema>
                          <GrausDeQualidade>
                                <AcuraciaConteudo>
                                        Informação coletada na emergência. Pode
                                        Conter erros.
                                </AcuraciaConteudo>
                                <Completeza>
                                        Não inclui informações de hospitais que não
                                        Possuem prontuário eletrônico.
                                </Completeza>
                                <Reputacao>Fonte: Hospitais</Reputação>
                          </GrausDeQualidade>
                    </root>
             </Value>
      </Cell>
      <Cell CellOrdinal="16">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="17">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="18">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="19">
             <Value xsi:type="xsd:double">80</Value>
      </Cell>
      <Cell CellOrdinal="20">
             <Value xsi:type="xsd:double">80</Value>
      </Cell>
      <Cell CellOrdinal="21">
             <Value xsi:type="xsd:double">80</Value>
      </Cell>
      <Cell CellOrdinal="22">
             <Value xsi:type="xsd:double">100</Value>
      </Cell>
      <Cell CellOrdinal="23">
             <Value xsi:type="xsd:string">
                    <root>
                           <xsd:schema xmlns:xsd=http://www.w3.org/2001/XMLSChema>
                          <!- O esquema XML da anotacao pode ser incluído aqui-->
                          </xsd:schema>
                          <GrausDeQualidade>
                                <Precisao>
                                        Valores estimados pelos profissionais dos
                                        serviços de orientação educacional nas
                                        escolas.
                                </Precisao>
                                <Completeza>
                                        Apenas informações de escolas estaduais.
                                </Completeza>
                                <Reputacao>Fonte:Secretaria de Educacao</Reputação>
                          </GrausDeQualidade>
                    </root>
             </Value>
      </Cell>
</CellData>
<!- Fim dos dados no formato MDDataSet -- >
```

```
</root>
</m:return>
</m:ExecuteResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Observações:

Durante a execução do método Execute, o Serviço de Qualidade traduz a consulta original para as seguintes consultas aos Cubos de Qualidade:

Consulta SQL (*Structured Query Language*) para obtenção do número do grupo em GRUPOS_DIMENSOES que contém apenas as dimensões envolvidas (considerando que o resultado seja armazanedo numa variável denominada GRUPO):

```
SELECT G1.PK_GR_DIMENSOES

FROM GRUPO_DIMENSOES G1, DIMENSOES D1

WHERE D1.NOME = 'MUNICIPIO' AND
D1.PK_DIMENSOES = G1.PK_DIMENSOES AND
G1.PK_GR_DIMENSOES NOT IN

(SELECT G2.PK_GR_DIMENSOES
FROM GRUPO_DIMENSOES G2, DIMENSOES D2
WHERE D2.NOME <> 'MUNICIPIO' AND
D2.PK_DIMENSOES = G2.PK_DIMENSOES)
```

Comando MDX para seleção dos graus de qualidade dos fatos envolvidos:

Comando MDX para seleção das anotações sobre os fatos envolvidos: