CONFIABILIDADE DE DADOS EM AMBIENTES DE BUSINESS INTELLIGENCE: UMA ABORDAGEM FUZZY BASEADA EM TAXONOMIAS DE PROBLEMAS DE QUALIDADE

WESLEY GONGORA DE ALMEIDA

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

CONFIABILIDADE DE DADOS EM AMBIENTES DE BUSINESS INTELLIGENCE: UMA ABORDAGEM FUZZY BASEADA EM TAXONOMIAS DE PROBLEMAS DE QUALIDADE

WESLEY GONGORA DE ALMEIDA

ORIENTADOR: RAFAEL TIMÓTEO DE SOUSA JÚNIOR

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: 481/2012
BRASÍLIA/DF: MAIO - 2012

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

CONFIABILIDADE DE DADOS EM AMBIENTES DE BUSINESS INTELLIGENCE: UMA ABORDAGEM FUZZY BASEADA EM TAXONOMIAS DE PROBLEMAS DE QUALIDADE

WESLEY GONGORA DE ALMEIDA

DISSERTAÇÃO DE MESTRADO ACADÊMICO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

APROVADA POR:

_____

Prof. Rafael Timóteo de Sousa Jr., Dr., ENE/UnB
(Orientador)

_____

Prof. Flávio Elias Gomes de Deus, Dr., ENE/UnB
(Examinador Interno)

_____

Prof. Georges Amvame Nze, Dr., FGA/UnB
(Examinador Externo)

BRASÍLIA/DF, 16 DE MARÇO DE 2012.

FICHA CATALOGRÁFICA

Almeida, Wesley Gongora de.
A447c Confiabilidade de dados em ambientes de Business Intelligence: uma abordagem fuzzy baseada em taxonomias de problemas de qualidade / Wesley Gongora de Almeida. -- 2012.
xvii, 90 f. : il. ; 30 cm.

Dissertação (mestrado) - Universidade de Brasília, Faculdade de Tecnologia, Departamento de Engenharia Elétrica, 2012.
Inclui bibliografia.
Orientação: Rafael Timóteo de Sousa Júnior.

1. Confiabilidade (Engenharia). 2. Armazenamento de Dados. 3. Inteligência competitiva (Administração). 4. Lógica difusa. I. Sousa Júnior, Rafael Timóteo de. II. Título.

CDU 658 . 012 .2

REFERÊNCIA BIBLIOGRÁFICA

ALMEIDA, W. G. (2012). Confiabilidade de Dados em Ambientes de Business Intelligence: Uma Abordagem Fuzzy Baseada em Taxonomias de Problemas de Qualidade, Dissertação de Mestrado em Engenharia Elétrica, Publicação 481/2012, Departamento de Engenharia Elétrica, Universidade de Brasília, DF, 90p.

CESSÃO DE DIREITOS

AUTOR: Wesley Gongora de Almeida

TÍTULO: Confiabilidade de Dados em Ambientes de Business Intelligence: Uma Abordagem Fuzzy Baseada em Taxonomias de Problemas de Qualidade.

GRAU: Mestre ANO: 2012

_____

Wesley Gongora de Almeida
QNN 06 Conjunto M Casa 09, Ceilândia
CEP: 72.220-073 - Brasília/DF - Brasil

Dedico este trabalho aos meus pais, Jacy e Odete, ao meu irmão Welder, a minha noiva Luana e, principalmente, a Deus.

Wesley Gongora de Almeida

AGRADECIMENTOS

Agradeço primeiramente a Deus por ser o meu guia em todos os momentos e ter permitido mais essa conquista.

A minha família, maiores professores, por terem me ensinado coisas que a escola não ensina. Sinceros agradecimentos aos meus pais, Jacy e Odete, e ao meu irmão, Welder.

A minha noiva Luana Soares Ferreira pelo carinho e companheirismo.

Agradeço aos amigos que me apoiaram e ajudaram direta e indiretamente nesta caminhada.

Ao meu professor e orientador, Rafael Timóteo de Sousa Jr. Qualquer agradecimento seria pouco perto das diversas oportunidades concedidas nesta Universidade. Meus sinceros agradecimentos pelas lições de vida e por todo aprendizado.

Aos professores Flávio Elias, Georges Amvame Nze e Laerte Peotta pelo apoio e aprendizado.

Aos colegas e amigos do LabRedes da UnB. Em especial a Adriana, ao Wandemberg, ao Diego, a Bia, ao Fábio, a Andréia e ao Valério.

Meus agradecimentos pela amizade e apoio a Karla, Lúcio, Vera, Ana e Evangelista e a todo pessoal da secretaria do ENE/UnB (e do churrasco da Xerox).

Muito obrigado a todos!

"Mantenha seus PENSAMENTOS positivos, porque seus pensamentos tornam-se suas palavras.
Mantenha suas PALAVRAS positivas, porque suas palavras tornam-se suas atitudes.
Mantenha suas ATITUDES positivas, porque suas atitudes tornam-se seus hábitos.
Mantenha seus HÁBITOS positivos, porque seus hábitos tornam-se seus valores.
Mantenha seus VALORES positivos, porque seus valores... Tornam-se seu DESTINO."

Mahatma Gandhi

RESUMO

CONFIABILIDADE DE DADOS EM AMBIENTES DE BUSINESS INTELLIGENCE: UMA ABORDAGEM FUZZY BASEADA EM TAXONOMIAS DE PROBLEMAS DE QUALIDADE

Autor: Wesley Gongora de Almeida

Orientador: Prof. Rafael Timóteo de Sousa Júnior, Depto. de Engenharia Elétrica / Universidade de Brasília

Programa de Pós-graduação em Engenharia Elétrica

Brasília, 16 de março de 2012

O impacto da má qualidade dos dados sobre a tomada de decisão, a confiança organizacional e a satisfação do cliente é bem conhecida. Ademais, fatores emergentes, tais como o aumento no volume dos dados, têm agravado o problema. Nas organizações atuais, sistemas de *Business Intelligence (BI)* têm oferecido suporte à gestão de negócios e se constituindo uma evolução natural e lógica dos Sistemas de Apoio a Decisão. Neste novo cenário, implementações de soluções de BI tem falhado devido a má qualidade dos dados. Supondo que é possível avaliar a qualidade dos dados com base em metadados, a questão principal, então, é: Como fornecer ao usuário informações relativas à qualidade dos dados? Atrelado a esta questão, encontra-se um segundo fator relevante: Durante muito tempo, preocupou-se com a qualidade dos dados sem levar em consideração a questão da confiança. Esta dissertação apresenta uma nova visão a respeito da qualidade e da confiança dos dados, porque, ao contrário do senso comum, a qualidade dos dados não é o único fator influenciando a confiabilidade dos dados e estes dois conceitos não são necessariamente correlacionados. Baixa qualidade pode ser confiável em algumas situações e dados de alta qualidade podem ter baixa confiança em outro contexto. Em nosso trabalho, a avaliação da confiabilidade dos dados em ambientes de BI é baseada em um conjunto de métricas, obtidas a partir de uma taxonomia dos problemas de qualidade. Para representar a incerteza da avaliação, lógica *fuzzy* é empregada como método de obtenção de uma pontuação global de confiabilidade. Por fim, a proposta desenvolvida é avaliada através de simulações, de forma a ilustrar sua eficácia e demonstrar um avanço em relação aos métodos estado-da-arte conhecidos da literatura.

Palavras-chave: Confiabilidade, Armazenamento de Dados, Inteligência Competitiva, Qualidade de Dados, Lógica Difusa.

ABSTRACT

TRUSTWORTHINESS OF DATA IN BUSINESS INTELLIGENCE ENVIRONMENTS: A FUZZY APPROACH BASED ON TAXONOMY OF QUALITY PROBLEMS

Author: Wesley Gongora de Almeida

Supervisor: Prof. Rafael Timóteo de Sousa Júnior, Depto. de Engenharia Elétrica / Universidade de Brasília

Programa de Pós-graduação em Engenharia Elétrica

Brasilia, March 16, 2012

The impact of poor data quality on decision making, organizational trust and customer satisfaction is well known. Furthermore, emerging factors, such as increasing the volume of data, have aggravated the problem. In today's organizations, Business Intelligence (BI) systems have offered support to business management and providing a natural and logical evolution of Decision Support Systems. In this new scenario, implementations of BI solutions have failed due to poor data quality. Assuming it is possible to assess the quality of data based on metadata, the main question then is: How to provide the user with information relating to data quality? Tied to this question lies a second relevant factor: For a long time, worried about the quality of data without taking into account the question of trust. This dissertation presents a new vision about the quality and trustworthiness of the data, because, contrary to common sense, data quality is not the only factor influencing the trustworthiness of data and these two concepts are not necessarily correlated. Low quality can be unreliable in some situations and high-quality data can have little confidence in another context. In our study, evaluating the trustworthiness of data in BI environments is based on a set of metrics, obtained from taxonomy of quality problems. To represent the uncertainty of the evaluation, fuzzy logic is employed as a method of obtaining an overall score of trustworthiness. Finally, the proposal developed is evaluated through simulations, in order to illustrate its effectiveness and demonstrate an improvement over methods state-of-the-art known from the literature.

Keywords: Trustworthiness, Data Warehousing, Business Intelligence, Data Quality, Fuzzy Logic.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

API	Interface de Programação de Aplicativos (do inglês *Application Program Interface*)
BI	Inteligência de Negócios (do inglês *Business Intelligence*)
CRM	Gestão de Relacionamento com o Cliente (do inglês *Customer Relationship Management*)
DSS	Sistema de Apoio a Decisão (do inglês *Decision Support System*)
DW	Armazém de Dados (do inglês *Data Warehouse*)
EIS	Sistema de Informação Executiva (do inglês *Executive Information System*)
ETL	Extração, Transformação e Carga (do inglês *Extraction, Transformation and Loading*)
GIS	Sistema de Informação Geográfica (do inglês *Geographic Information System*)
GRPU	Gerência Regional do Patrimônio da União
MP	Ministério do Planejamento, Orçamento e Gestão
QD	Qualidade de Dados
QI	Qualidade da Informação
ODS	Armazém de Dados Operacionais (do inglês *Operational Data Store*)
OLAP	Processamento Analítico em Tempo Real (do inglês *Online Analytical Processing*)
OLTP	Processamento de Transações em Tempo Real (do inglês *Online Transactional Processing*)
SAW	Ponderação Aditiva Simples (do inglês *Simple Additive Weighting*)
SI	Sistema de Informação
SPU	Secretaria de Patrimônio da União
TI	Tecnologia da Informação
XML	Linguagem de Marcação Estendida (do inglês *Extensible Markup Language*)

1. INTRODUÇÃO

Nesta dissertação abordamos o problema da avaliação da confiabilidade dos dados em ambientes de *Data Warehouse/Business Intelligence*, de uma maneira que seja tão completa e objetiva quanto possível. Para realizarmos nosso objetivo, um conjunto de técnicas e um novo método de avaliação são necessários. Este trabalho descreve o conjunto de procedimentos utilizados na construção da solução e os resultados obtidos.

1.1. DESCRIÇÃO DO PROBLEMA

A questão da Qualidade de Dados (QD) é tão antiga quanto à dos dados em si. O impacto da má qualidade dos dados sobre a tomada de decisão, a confiança organizacional e a satisfação do cliente é bem conhecida. De acordo com [65] "*o custo total da má QD é entre 8% e 12% das receitas das empresas*". Além disso, se os dados forem de má qualidade, isso se refletirá nos resultados produzidos (princípio "lixo entra, lixo sai") [85]. Dados, como os disponíveis em um sistema de informação, são sempre de alguma forma imperfeitos [19].

Nas organizações atuais, sistemas de *Data Warehouse* (DW) têm assumido um papel cada vez mais relevante, em virtude de constituírem-se uma evolução natural e lógica dos Sistemas de Apoio a Decisão (SAD). Problemas de qualidade nestes sistemas constituem-se uma grande preocupação, já que um DW com dados sujos pode ser prejudicial e, na melhor das hipóteses, não ser confiável [53]. Problemas de qualidade como estes podem, portanto, ter impactos significativamente negativos sobre a eficiência de uma organização, enquanto dados de alta qualidade são freqüentemente cruciais para o sucesso de uma empresa [28, 83].

Em [57], é indicado que 41% dos projetos de *Data Warehouse* falham, principalmente, devido à QD insuficiente. Em um estudo anterior, foi observado que 67% dos gerentes de *marketing* pensam que a satisfação de seus clientes sofre de má QD [66]. Esses números, apesar de pertencerem a trabalhos pioneiros, seguem, nos dias atuais, a ilustrar os desafios no campo de pesquisa da qualidade.

A área de estudo em questão tem ganhado cada vez mais importância, na teoria e na prática, devido a vários fatores emergentes. Este cenário tem elevado o problema a um nível estratégico e os riscos para a tomada de decisão têm aumentado. Primeiro, há claras

implicações que se relacionam com o grande volume de dados produzidos pelas organizações de hoje. Em segundo lugar, têm-se visto nos últimos anos um aumento na volatilidade e na diversidade de fontes de dados. Essa diversidade refere-se à arquitetura dos dados, apresentando-se estruturados, semi-estruturados e não estruturados, tais como dados multimídia em formato de vídeo, mapas, imagens, etc. [1].

Durante os últimos anos, algumas técnicas têm sido aplicadas para tratar do problema. O uso de metadados de qualidade que reflitam a linhagem dos dados tem sido bastante empregado para o monitoramento da qualidade [41, 42, 43, 44, 74]. Entretanto, é preciso identificar os componentes relevantes corretamente. Um repositório de metadados relevante é tipicamente composto de: (a) fontes de dados, (b) serviços de transformação / processo de ETL e (c) requisitos do usuário [5, 40].

Supondo que é possível avaliar a qualidade dos dados com base em metadados, as duas questões principais, então, são:

- Como posso saber se as informações são confiáveis?
- Como fornecer ao usuário essas informações, juntamente com os resultados da consulta?[1]

Em várias discussões relacionadas com metadados de QD e modelagem, tornou-se evidente que a confiabilidade desempenha um importante papel na qualidade dos dados. Assim, o primeiro passo desta investigação parte com o objetivo de explorar a relação entre a confiabilidade e a qualidade dos dados, no contexto de diferentes arquiteturas de gerenciamento de dados. Baseado na noção de dados em um sistema de informação, e na comprovação das informações representadas pelos dados, observou-se que várias noções de confiança dos dados podem ser definidas.

Em particular, um, então pode distinguir entre a noção de "fato" e de "crença" [5]. Se as informações representadas pelos dados podem ser verificadas por algum meio, considera-se um fato. A verificação pode ser baseada em teorias estabelecidas ou com base na "confiança" em uma autoridade para avaliar as informações. Se as informações representadas pelos dados não podem ser verificadas, então não é certa a "crença" naquele pedaço de informação.

[1] Questionamentos semelhantes a este já foram feitos em outros trabalhos, e.g., Gertz et al., 2004 [5]. No entanto, como será visto nesta dissertação, muitas lacunas ainda se encontram presentes.

A crença pode ser criada com base em alguns indícios do comportamento passado ou por outros meios. Isto é capturado pela noção de "reputação", que é a memória e o resumo do comportamento baseado em transações passadas. Assim, a confiança que se fundamenta na crença baseada em evidências, ou na reputação, é um componente mensurável e objetivo de confiança.

1.2. OBJETIVOS E CONTRIBUIÇÕES

O trabalho desenvolvido e apresentado nesta dissertação não tem a pretensão de fornecer soluções definitivas, mas apenas contribuir com uma pequena parcela de conhecimento, a fim de permitir algum avanço técnico-científico na área de estudo em questão. Motivado por pesquisas anteriores, e.g., Cappiello et al. [4] e pela necessidade de estabelecer um modelo de confiabilidade de dados para aplicações de BI apoiados por repositórios de DW, as principais contribuições deste trabalho são:

- Estabelecimento de uma relação entre a confiabilidade e a qualidade dos dados, no contexto das aplicações de BI;
- Identificação das dimensões que refletem a noção da confiabilidade dos dados;
- Definição de uma taxonomia dos problemas de QD para ambientes de DW/BI;
- Estabelecimento de métricas a partir de taxonomias de problemas de QD para ambientes de DW/BI;
- Definição de um modelo *fuzzy* para a avaliação da confiabilidade dos dados;
- Definição de uma arquitetura de BI orientada a confiabilidade dos dados, capaz de fornecer informações relativas à qualidade dos dados, juntamente com os resultados da consulta. A solução deve abranger todos os tipos de saídas *(outputs)* disponíveis em interfaces de BI apresentados por [2].

1.3. APRESENTAÇÃO DA DISSERTAÇÃO

Esta dissertação está organizada em seis capítulos, incluindo esta introdução.

Capítulo 2 – Sistemas de Business Intelligence: Neste capítulo apresentamos um referencial teórico sobre *Business Intelligence*. Após sua conceituação, a descrição de cada componente da arquitetura de BI é realizada.

Capítulo 3 – Confiabilidade dos Dados: Neste capítulo, é dada uma visão geral a respeito das pesquisas relacionadas aos temas tratados neste trabalho. Inicialmente, a questão da qualidade dos dados em ambientes de BI é introduzida. Em seguida, a perspectiva da confiabilidade dos dados é analisada. Por fim, apresentamos uma revisão a respeito das taxonomias dos problemas de QD e de métricas globais para avaliar a confiabilidade dos dados.

Capítulo 4 – Modelo *Fuzzy* de Confiabilidade dos Dados para Ambientes de BI: Neste capítulo, a partir de um modelo matemático probabilístico, estendemos o conceito da qualidade de dados através da perspectiva de dados confiáveis. No final do capítulo, são demonstradas as vantagens desta abordagem em relação aos métodos disponíveis na literatura.

Capítulo 5 – Simulações e Análise dos Resultados: Na primeira parte do capítulo apresentamos as simulações realizadas e os resultados obtidos. A segunda etapa consiste de um estudo de caso para demonstrar à viabilidade de implementação da proposta.

Em seguida, o Capítulo 6 apresenta as conclusões e discute sugestões de continuidade para este trabalho.

2. SISTEMAS DE BUSINESS INTELLIGENCE

Este capítulo inicia-se com algumas definições fundamentais a respeito do tema *Business Intelligence* (BI) na Seção 2.1. Posteriormente, a arquitetura típica de um Sistema de BI é apresentada e detalhada através da descrição de cada um de seus componentes na Subseção 2.1.1.

2.1. SISTEMAS DE BUSINESS INTELLIGENCE

O termo inglês *Business Intelligence* (BI), traduzido como Inteligência de Negócios, refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios. Constitui-se como uma evolução natural e lógica dos Sistemas de Apoio a Decisão e, em última instância, do próprio conceito de Tecnologia da Informação.

Conforme definido por Angel em [11], BI pode ser pensado como *"extração e análise de informações relevantes, tornando-as acessíveis para apoio no processo de tomada de decisão"* [24] . BI envolve a captura de informações de muitos outros sistemas, tais como Ferramentas de Processamento Analítico *Online* (OLAP), Mineração de Dados *(Data Mining)*, Sistemas de Apoio à Decisão *(Decision Support System - DSS*), Sistemas de Informação Geográfica *(Geographic Information System - GIS)*, entre outros [12]. A Figura 2.1 mostra alguns dos Sistemas de Informação (SI) que são utilizados pelos Sistemas de BI [12]. Adaptado de (Negash, 2008).

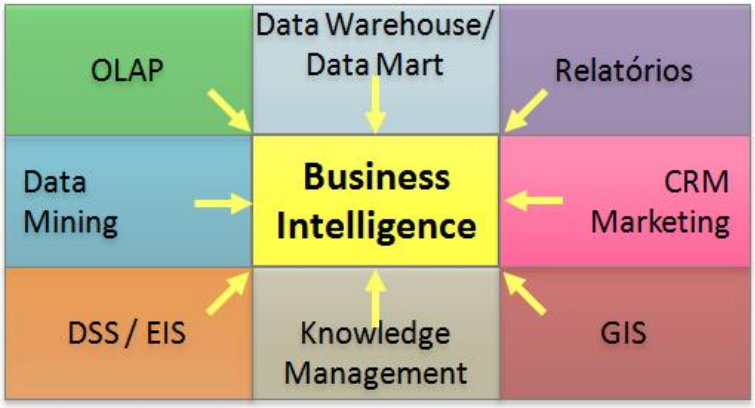


Figura 2.1: Relação do BI com outros SI. Adaptado de (Negash, 2008)

De acordo com [16], *Business Intelligence* surgiu para solucionar o maior problema das organizações, que é a ausência de conhecimento adequado sobre si mesmo, ou seja,

sobre seu conjunto de informações. BI serve para eliminar as dúvidas e a ignorância das organizações sobre seus dados, usando, para isso, mecanismos capazes de acessar dados e explorar as informações, analisando-as e desenvolvendo percepções e entendimentos a seu respeito, permitindo incrementar e tornar a tomada de decisão pautada em informações mais consistentes.

O conceito foi formalizado na década de 80, mas seus princípios básicos já eram usados pelos povos antigos. Há indícios que sociedades do Oriente Médio antigo, séculos antes de Cristo, utilizavam os princípios básicos do BI quando cruzavam informações obtidas junto à natureza em benefício de suas aldeias, e.g., o comportamento das marés, os períodos chuvosos e de seca, a posição dos astros, etc.

2.1.1. Arquitetura de um Sistema de BI

A arquitetura de um sistema de BI é altamente distribuída. Neste contexto, é imprescindível a utilização de mecanismos que garantam a interoperabilidade e a segurança. A Figura 2.2 apresenta a arquitetura típica de um sistema de BI [2, 18, 21].

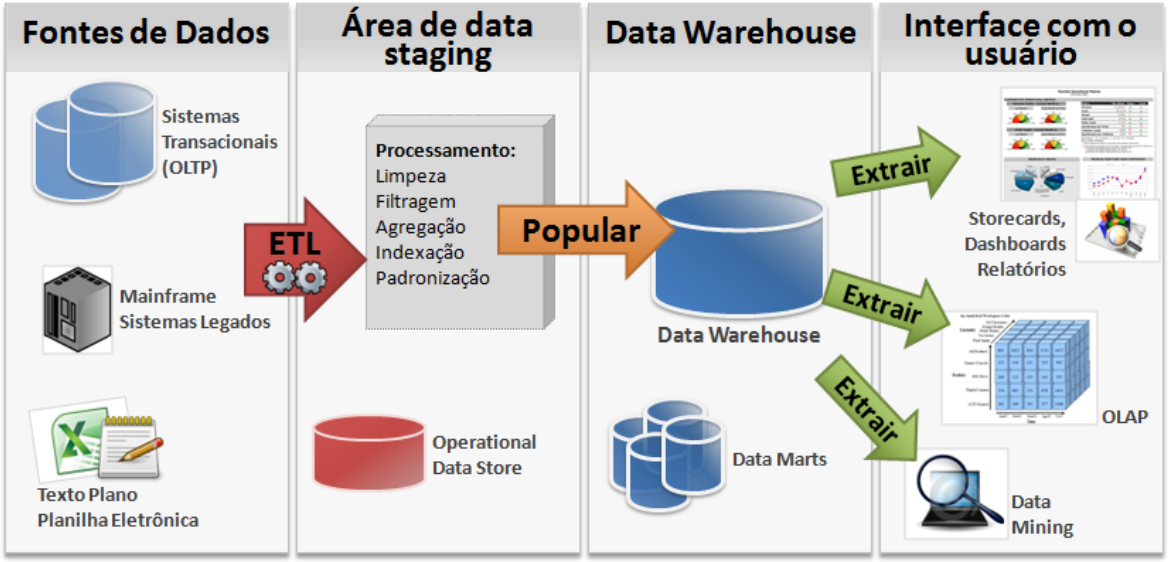


Figura 2.2: Arquitetura típica de um Sistema de BI

2.1.1.1. Fontes de Dados

Constitui-se como a origem dos dados pelo qual o ambiente de BI é alimentado. Fontes de dados tradicionais são geralmente compostas de sistemas de Processamento de Transações em Tempo Real (do inglês *Online Transactional Processing* - OLTP). Entretanto, uma grande diversidade de fontes de dados pode ser utilizada. Recentemente,

observa-se um interesse crescente pela adoção de fontes de dados semi-estruturadas e não estruturadas, tais como dados multimídia, i.e., vídeo, mapas, imagens, etc. [1].

2.1.1.2. Área de Data Staging

As empresas gastam bilhões de dólares na obtenção de dados limpos e inequívocos em seu DW. Kimball et al. [43] observa que mais de 70% do tempo e esforço gasto para a execução de um projeto de *Data Warehouse* é consumido nos processos de tratamento dos dados. A área de *Data Staging* é, portanto, tanto uma área de armazenamento como um conjunto de processos, e normalmente denomina-se ETL (do inglês *Extract, Transform and Load*). Esta sigla é uma alusão as três fases principais de manipulação das bases de dados: Extração, Transformação e Carga.

A extração é a fase responsável pela coleta dos dados de diversas fontes internas e/ou externas à organização. A transformação unifica os formatos e faz a "limpeza" dos registros incompletos e das inconsistências dos dados para atender às necessidades de negócios [23]. Durante esta etapa, os dados eventualmente poderão ser armazenados em um repositório intermediário entre as fontes e o DW, conhecido como ODS, do inglês *Operational Data Store*. Por fim, os dados tratados são então carregados em um repositório de dados, tipicamente um *Data Warehouse* ou *Data Mart*.

2.1.1.3. Data Warehouse: Estrutura de armazenamento de informações

Um sistema de DW é caracterizado como um banco de dados multidimensional que armazena dados orientados por assunto, integrado, variável em relação ao tempo e não volátil, e que muitas vezes é modelado através de um esquema estrela composto de tabelas Fato e Dimensão [43].

De acordo com Kimball et al. [43], o DW fornece acesso a dados corporativos ou organizacionais, seus dados são consistentes podendo ser separados e combinados usando-se qualquer medição possível no negócio. Alguns autores defendem que o DW não consiste apenas em dados, mas em um conjunto de ferramentas para consultar, analisar e apresentar informações [45], incorporando o próprio conceito do BI. De qualquer forma, um DW/BI deve ser um local onde se publicam dados confiáveis, sendo a qualidade um desses impulsos à reengenharia de negócios.

Os *Data Marts* são repositórios de dados semelhantes aos DWs. Diferenciam-se por focalizarem uma ou mais áreas específicas na organização, tornando-se subconjuntos de dados de um DW.

2.1.1.4. Interface com o usuário (análises*)*

São as ferramentas de análise de dados, com a possibilidade de descoberta de informações explícitas e implícitas, úteis para as organizações. Podem ser classificadas como ferramentas: de processamento analítico em tempo real (*Online Analytical Processing* - OLAP), de análise exploratória de dados (AED), e.g., *Dashboards e Storecards*, e de processo de descoberta de conhecimento em bancos de dados (*Knowledge Discovery in Databases* - KDD), e.g., Mineração de Dados.

Nas técnicas OLAP, a expressão "tempo real" significa que as operações devem ter uma resposta imediata. Já o termo "analítico" demonstra o uso de teorias analíticas para tornar as buscas possíveis. A palavra "processamento" reforça a característica de intenso processamento na utilização de uma grande quantidade de dados [15].

3. CONFIABILIDADE DOS DADOS

A qualidade dos dados em Sistemas de Informação tem sido um campo de pesquisa ativo nos últimos 20 anos. É possível encontrar na literatura diversas propostas sendo elaboradas para os mais variados domínios. Entretanto, o problema da qualidade continua longe de ser resolvido [1] por diversos motivos, como veremos adiante.

Os temas gerais investigados nesta dissertação vêm sendo estudados por diversos autores ao longo dos últimos anos na literatura. Os tópicos mais específicos, referentes à confiabilidade dos dados, também são examinados em alguns estudos mais recentes, porém, com abordagens diferentes. Neste capítulo, é dada uma visão geral a respeito das pesquisas relacionadas aos temas tratados neste trabalho.

O capítulo está organizado da seguinte maneira. A Seção 3.1 apresenta os conceitos relativos à qualidade de dados. Esta conceituação é, em seguida, estendida para abordar a cobertura da qualidade através de um resumo das metodologias citadas na literatura para manutenção da qualidade, na Seção 3.2. Em seguida, na Seção 3.3, a questão da confiança dos dados é apresentada. Por fim, nas Seções 3.4 e 3.5, apresentamos uma revisão a respeito de taxonomias da confiança e de métricas globais para avaliar a confiabilidade dos dados.

3.1. QUALIDADE DE DADOS

Problemas de QD podem ter efeitos terríveis sobre um negócio. Conforme defendido em [9], a *"qualidade dos dados é um grande inibidor para o sucesso dos projetos de BI, podendo causar desconfiança do usuário e o conseqüente abandono do sistema"*. A conseqüência de dados ruins é sentida pela corporação através das más decisões estratégicas.

O problema da qualidade de dados é estudado em diversas perspectivas. Por exemplo, em [13] o problema da inconsistência das várias fontes em relação a preferência do usuário é tratada na consulta ao banco de dados (consultas em linguagem conhecida como *SQL*). Em [7] as dimensões relevantes da QD são organizadas segundo categorias de QD (como intrínseca, contextual, de acessibilidade e de representação). Uma lista mais extensa dos aspectos da qualidade dos dados é sugerida por diversos consumidores de dados e pode ser encontrada em [8], onde 179 dimensões de qualidade são citadas em uma

pesquisa junto a 137 pessoas, compostas por profissionais da indústria e por estudantes de MBA de uma grande universidade dos EUA.

Revisões de literatura mais recentes revelam que a preocupação com a qualidade dos dados está presente sob diversas perspectivas diferentes. Sadiq et al. [1] em seu levantamento sobre os 20 anos de pesquisas em qualidade de dados, em 2011, ratifica a expressiva quantidade de artigos sobre o tema durante o período de 1990-2009: 31.701 artigos. Entretanto, a questão da qualidade de dados está longe de ser considerada uma questão resolvida, conforme argumenta o próprio Sadiq et al. [1].

Mais do que nunca, empresas, governos e organizações de pesquisa se baseiam no intercâmbio e compartilhamento de dados. É amplamente reconhecido que lidar com problemas de QD é caro e demorado, levando a ramos de novas tecnologias de TI que focam exclusivamente a avaliação, a manutenção e a limpeza dos dados em uma organização [5].

A pesquisa sobre a QD começou no contexto de sistemas de informação [7, 37] e foi estendida para diversos outros contextos, e.g., sistemas colaborativos, armazéns de dados, *e-commerce* e em Portais Web, devido às características particulares pertinentes a cada domínio [30, 34]. Alguns trabalhos na área em questão, por outro lado, analisaram a qualidade na perspectiva dos consumidores de dados [8, 31, 35].

Conforme afirma Wand e Wang (1996), *"a qualidade dos dados depende dos processos de concepção e produção envolvidos na geração dos dados. Para projetar uma melhor qualidade, é necessário primeiro entender o que significa qualidade e como ela é medida"* [10].

3.1.1. Conceituando Qualidade de Dados

Pode-se, provavelmente, encontrar tantas definições para dados de qualidade quanto se pode imaginar, pois há inúmeros trabalhos sobre o assunto. Conforme afirma Beverly et al. [9], qualidade da informação (ou qualidade dos dados) é *"uma ciência inexata, em termos de avaliações e benchmarks"*. Muitas vezes, dados de alta qualidade são simplesmente *"dados que estão aptos para serem utilizados pelos consumidores de dados"* [3, 5].

Na literatura pertinente, o conceito de Qualidade de Dados é freqüentemente definido como *"adequação ao uso"*, ou seja, a capacidade de uma coleta de dados atender à necessidade do usuário [7, 31].

Outro ponto a ser levado em consideração é que os termos "dado" e "informação" são freqüentemente utilizados como sinônimos [30, 32], como será feito também no escopo desta dissertação.

3.1.2. Problemas de Qualidade de Dados

Atualmente, a importância e a utilidade dos dados são largamente reconhecidas. É sabido que os dados são um ativo chave para qualquer ambiente organizacional manter-se competitivo. No entanto, é sabido também que os dados encontram-se afetados por diversos problemas de qualidade [28]. Entre outros problemas, os dados podem conter: violações de domínio, sinônimos, violações de restrição de integridade, registros duplicados, violações à integridade referencial, etc. [66]. Diante deste cenário, a existência de dados com qualidade não é regra, mas sim exceção, em algumas fontes de dados.

Como a identificação dos problemas de QD é freqüentemente desenvolvida em uma base *ad hoc* para resolver problemas específicos e práticos, muitas lacunas encontram-se manifestas [40]. Somente através de investigações sistemáticas os problemas de QD podem ser mais bem tratados. Para atender a esse quesito, apresentamos em seguida uma revisão das taxonomias de problemas de QD existentes na literatura.

Antes de prosseguirmos, é importante que o conceito taxonomia seja bem definido, já que ele é bastante utilizado, e em diferentes contextos. A palavra Taxonomia vem do grego *τασσεῖν* ou *taxis* = "classificar, organizar" e *νόμος* ou *nomos* = "lei, ciência, administrar" e pode ser definido como a ciência de classificação, denominação e organização de um sistema pré-determinado e que tem como resultante um *framework* conceitual para discussões, análises e/ou recuperação de informação.

Na literatura, cinco taxonomias de problemas de QD são propostas. A forma como os problemas encontram-se organizados em cada uma destas taxonomias é apresentada a seguir. Inspirados na revisão de taxonomias dos problemas de qualidade de [66], apresentamos uma versão atualizada do quadro comparativo entre as propostas, conforme a seguir:

- **Taxonomia de Rahm e Do (Rahm e Do, 2000)** – Nesta taxonomia [67] é feita uma distinção entre problemas de QD mono-fonte e multi-fonte de dados. Os problemas mono-fonte e multi-fonte encontram-se divididos nos que se relacionam com o esquema dos dados e nos que se relacionam com as instâncias. Os problemas relacionados com o esquema podem ser evitados melhorando o seu desenho, a sua transformação ou a sua integração. Os problemas nas instâncias correspondem a erros e inconsistências nos dados que não podem ser evitados através do esquema. Nos problemas mono-fonte é efetuada uma distinção entre os problemas que ocorrem no: (i) atributo (e.g., valor em falta, erro ortográfico); (ii) registro (e.g., violação de dependência funcional); (iii) tipo de registro (e.g., registros duplicados); e, (iv) fonte de dados (e.g., referência errada). Não são fornecidas informações sobre a abordagem usada na identificação dos problemas.

- **Taxonomia de Müller e Freytag (Müller e Freytag, 2003)** – Nesta taxonomia [68], os problemas de QD são classificados genericamente com sendo sintáticos, semânticos e de cobertura. Os problemas sintáticos dizem respeito aos aspectos relacionados com sintaxe e valores usados na representação das entidades (e.g., violações de sintaxe). Os problemas semânticos impedem os dados de constituírem uma representação não redundante e não ambígua do mundo real (e.g., registros duplicados). Os problemas de cobertura estão relacionados com a quantidade de entidades e propriedades das entidades do mundo real que, de fato, se encontram armazenadas na tabela (e.g., valores em falta). Nesta taxonomia, além de muito genéricos, os problemas de QD encontram-se limitados aos que ocorrem ao nível de uma só tabela, daí que muitos outros estejam em falta. Nada é dito sobre a forma como se chegou aos problemas apresentados em cada grupo (i.e., problemas sintáticos, problemas semânticos e problemas de cobertura)

- **Taxonomia de Dados Sujos (Kim et al., 2003)** – Nesta taxonomia [53] é apresentada uma relação bastante completa de problemas de QD, sendo descrita a lógica subjacente à sua elaboração. Pode ser considerada a mais completa, sob a perspectiva quantitativa, se comparada com todas as demais citadas na literatura. Nela, seus autores adotam uma abordagem hierárquica descendente, aumentando sucessivamente o grau de detalhe dos problemas. O ponto de

partida consiste em uma classificação genérica dos problemas de QD em três grandes classes: (i) valores em falta; (ii) valores existentes, mas errados (e.g., violação de integridade referencial, violação de unicidade); e, (iii) valores existentes e corretos mas não utilizáveis (e.g., utilização de abreviaturas, unidades de medidas diferentes). Este último caso advém de não ter sido adotado um padrão de representação comum na introdução dos valores. A taxonomia surge da decomposição hierárquica destas três classes genéricas de manifestação de problemas de QD.

- **Taxonomia dos Problemas de QD (Oliveira et al., 2005) –** Nesta taxonomia [54], uma relação abrangente de 35 problemas de QD é proposta. Os problemas encontram-se organizados pelo nível de granularidade em que ocorrem. A aproximação adotada foi ascendente, i.e., começou-se por analisar o nível mais elementar e terminou-e no mais complexo: (i) problemas ao nível do atributo; (ii) problemas ao nível da tupla; (iii) problemas ao nível da relação; e (iv) problemas ao nível de múltiplas relações/múltiplas fontes de dados. Um método para detectar problemas de QD como árvores binárias é também proposto para cada nível de abstração. Cada um dos problemas foi reapresentado em [66] através de definições formais, eliminando, assim, possíveis subjetividades típicas das definições textuais.

- **Taxonomia dos Dados Sujos baseados em Regras (Li et al., 2011) –** Nesta taxonomia, publicada mais recentemente [56], os problemas de dados sujos de [53] são revisados e classificados em 37 problemas de QD. Se comparado com as demais obras, esta taxonomia fornece uma coleção de problemas de QD mais abrangente e mais próxima dos problemas de QD em DW/BI. Em seguida, o trabalho apresenta um método para lidar com problemas de QD através da seleção de dados sujos a serem tratados prioritariamente (ou seja, baseado em regras). Dentro deste método, os autores propõem uma classificação dos problemas de QD baseada em dimensões da qualidade.

Quando se comparam os problemas que constituem cada uma das taxonomias, constata-se que estes são diferentes. As diferenças não residem somente ao nível dos termos usados para denominar cada problema de QD. Apesar de existir um conjunto de problemas comuns, i.e., pertencendo a mais de uma taxonomia, também há problemas que

surgem em apenas uma destas (e.g. *Tuplas ausentes* apenas é mencionado em [68]). Este fato indica que nenhuma das cinco taxonomias pode ser considerada completa, i.e., engloba todos os problemas de qualidade que afetam os dados. A reunião dos problemas que constam das cinco taxonomias permite obter um conjunto bastante vasto de problemas de QD, mas que, ainda sim, pode não ser completo. Não é fácil detectar eventuais problemas que possam estar em falta neste conjunto quando a sua identificação não resulta de um método sistemático, mas sim de uma compilação de problemas. Neste propósito, as taxonomias que explicitam claramente o método usado na identificação dos problemas são a de Kim et al. [53], de Oliveira et al. [54] e a de Li et al. [56].

Uma vez que todas as taxonomias existentes na literatura são destinadas a bancos de dados transacionais, alguns problemas de qualidade manifestos em *Data Warehouses* não se encontram identificados. Em segundo lugar, muitos problemas, quando aplicados em ambientes de DW, apresentam-se parcialmente mapeados, enquanto outros devem ser desconsiderados. Uma vez que neste trabalho pretende-se conceber e desenvolver um modelo de taxonomia que dê suporte aos problemas de QD em ambientes de DW, a sua identificação exaustiva reveste-se de uma especial importância. Estes motivos levaram à elaboração de uma nova taxonomia que ofereça uma cobertura adequada aos problemas de QD presentes nos modelos de dados multidimensionais utilizados em DW, cuja apresentação será realizada no Capítulo 4.

3.2. METODOLOGIAS RELACIONADAS À QD

Encontram-se na literatura muitas pesquisas e diferentes enfoques com o propósito de prover a manutenção da QD. A Tabela 3.1 relaciona as principais metodologias existentes [28]. É apresentado ainda, na última coluna, a identificação dos trabalhos no qual abrangem medições de qualidade (representados por **+**).

Tabela 3.1: Metodologias para a qualidade de dados. Adaptado de (Batini et al., 2009)

Abreviatura	Nome por extenso	Referência (conforme [28])	Medição da qualidade
GQM	Goal Question Metric	Basili, 1994	+
TDQM	Total Data Quality Management	Wang, 1998 [32]	+
DWQ	The Data Warehouse Quality Methodology	Jeusfeld et al., 1998	+
TIQM	Total Information Quality Management	English, 1999	+

(continua)

Abreviatura	Nome por extenso *(continuação)*	Referência (conforme [28])	Medição da qualidade
AIMQ	A methodology for information quality assessment	Lee et al. 2002 [33]	+
CIHI	Canadian Institute for Health Information methodology	Long and Seko, 2005	-
DQA	Data Quality Assessment	Pipino et al. 2002 [40]	+
IQM	Information Quality Measurement	Eppler and Munzenmaier 2002	+
ISTAT	ISTAT methodology	Falorsi et al 2003	+
AMEQ	Activity-based Measuring and Evaluating of product information Quality methodology	Su and Jin 2004	+
COLDQ	Loshin Methodology (Cost-effect of Low Data Quality	Loshin, 2004	+
DaQuinCIS	Data Quality in Cooperative Information Systems	Scannapieco et al. 2004	+
QAFD	Methodology for the Quality Assessment of Financial Data	De Amicis and Batini, 2004	+
CDQ	Comprehensive methodology for Data Quality management	Batini and Scannapieco, 2006	+

+ *Metodologia na qual é abrangido medições de QD*
- *Ausência de medições de QD*

A preocupação com medidas de qualidade sempre esteve presente nas metodologias de qualidade. Das propostas apresentadas na Tabela 3.1, apenas uma (i.e., a CIHI - *Canadian Institute for Health Information methodology*) não possui essa cobertura. A seguir apresentamos as considerações sobre as três abordagens mais difundidas, tanto no meio acadêmico-científico, quanto em ambientes corporativos (i.e. as metodologias GQM, TDQM e o DWQ):

- **GQM (*Goal Question Metric*):** Baseia-se na suposição de que, para se medir de maneira eficaz, alguns objetivos devem ser estabelecidos para que estes sirvam de rota para o estabelecimento de questões, que irão orientar a definição de métricas em um contexto particular. No GQM, os objetivos não devem ser avaliados diretamente, mas através de questionamentos que precisam ser respondidos durante a avaliação da qualidade.
- **TDQM (*Total Data Quality Management*):** A TDQM [32] é uma das metodologias mais importantes para o gerenciamento de QD. Muito utilizada como suporte para iniciativas de reengenharia de dados nas organizações, ela foi desenvolvida pelo *Massachusetts Institute of Technology* (MIT) e adota a

perspectiva da informação como um produto. A TDQM prevê métricas de qualidade da informação, mede a qualidade ao longo do ciclo de vida da informação, analisa e identifica as causas que geram problemas de qualidade e define a implementação do processo de melhoria da qualidade dos dados. Esta metodologia é um processo iterativo e incremental, onde quatro fases são identificas: definir, analisar, medir e melhorar.

- **DWQ (*Data Warehouse Quality*):** Estuda as relações entre objetos de qualidade e opções de modelagem no DW. Esta metodologia considera conceitos subjetivos de qualidade e provê uma classificação de metas de qualidade conforme definição do grupo de *stakeholders* que procuram essas metas. Eles consideram a diversidade de metas de qualidade existentes e utilizam metadados para defini-las.

Para que seja possível, então, um maior êxito nos projetos de BI e a obtenção de máxima QD na organização, a adoção de um processo de Governança de Dados é necessária.

3.2.1. Dimensões de Qualidade de Dados

Diversos critérios (ou dimensões[2]) da qualidade de dados podem ser considerados para a composição da opinião sobre dados de qualidade (ver Seção 3.3.2). Entretanto, algumas dimensões são consideradas fundamentais e amplamente citadas na literatura [8, 27, 28, 40, 41]:

- **Acurácia:** Representa a distância entre dois valores v e v', sendo v considerado o valor correto. Pode ser de nível sintático, semântico ou de conteúdo.
- **Credibilidade (ou confiabilidade):** São aceitos ou considerados como verdadeiros e autênticos.
- **Completude:** É o grau no qual os elementos de um esquema estão presentes nas instâncias. A suficiência dos dados para serem utilizadas na resolução de uma determinada tarefa.

[2] Na literatura, o termo "dimensão da qualidade" é geralmente utilizado para referenciar grupos de atributos da qualidade, embora o termo "critério da qualidade" também possa ser utilizado com esse significado.

- **Temporalidade:** Disponibilidade dos dados no tempo esperado, de acordo com os requisitos de tempo especificados pelo destino.
- **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. É definido, por alguns autores, como uma subcategoria da acurácia.
- **Ausência de dados duplicados:** Cada item de informação tem um único significado.
- **Facilidade de acesso:** Estão disponíveis ou são facilmente recuperados.

3.2.2. Medição da qualidade dos dados

Na literatura, as diferentes perspectivas sobre a medição da qualidade têm sido classificadas em duas perspectivas [57]: qualidade do modelo e qualidade de conformidade.

Qualidade de modelo denota o grau de correspondência entre as necessidades dos usuários e a especificação do sistema de informação (e.g., especificado por meio de esquemas de dados).

Em contraste, qualidade de conformidade representa o grau de correspondência entre a especificação e os cadastros existentes nos sistemas de informação (e.g., dados de esquemas *versus* conjunto de dados de clientes armazenados).

A distinção entre qualidade de modelo e qualidade de conformidade é importante dentro do contexto de quantificar a qualidade e/ou confiabilidade dos dados. Ela separa a análise, essencialmente subjetiva, da correspondência entre os requisitos dos usuários e os esquemas de dados especificados – do que é mais objetivo – da correspondência entre os esquemas de dados especificados e os valores de dados existentes. Na proposta apresentada neste trabalho, ambos os aspectos devem ser levados em consideração. Como veremos a seguir, o conceito da confiança aplicada em ambientes de BI engloba em sua natureza as duas perspectivas.

A avaliação da qualidade é uma tarefa difícil, devido às seguintes razões. Primeiramente, muitos critérios de QD são subjetivos e, portanto, não podem ser avaliados automaticamente (e.g., temporalidade, volume de dados, valor agregado, reputação das fontes). Em segundo lugar, muitas fontes não publicam metadados relacionados à qualidade. Em terceiro lugar, para fontes com grandes quantidades de dados, a avaliação da QD é geralmente desencorajada, sendo normalmente realizada com uma amostra dos dados, resultando em uma baixa precisão nos escores de qualidade.

Além dessas noções fundamentais, é preciso estar ciente das duas dimensões da dinâmica da QD: a visão passiva e a ativa [5, 28, 32]. A visão passiva pode ser caracterizada como o rastreamento e a observação dos dados relevantes e dos componentes de *software* ao longo do tempo. Esta visão está intimamente relacionada com a detecção de mudanças (nosso objeto de estudo, concomitante com outras abordagens [5]).

Os resultados da segunda dinâmica estão preocupados em influenciar a QD, resultando na imposição de mudanças nos dados e nos componentes. Assim, uma visão ativa visa impor algum tipo de ciclo de vida na QD [5, 28, 32].

No âmbito dos modelos e técnicas de medição empregadas, é possível encontrar diversos métodos, tais como medidas de dispersão, regressão linear, probabilidades condicionais, redes neurais, árvores de decisão, entre outras [28, 69]. Levando em consideração a subjetividade intrínseca do ponto de vista do consumidor de dados e a incerteza inerente à percepção de qualidade [37], Caro et al. [30] decidiu-se empregar uma abordagem probabilística (baseado em redes bayesianas e lógica *fuzzy*), proposto em [38] para avaliar a qualidade dos dados em um Portal *Web*. Sua metodologia é conhecida como PDQM. Este reafirma a idéia de que a QD precisa ser avaliada dentro do contexto de geração de dados [36].

3.3. A QUESTÃO DA CONFIANÇA EM AMBIENTES DE BI

A confiança e a reputação são conceitos estudados em diversas áreas, tais como economia, sociologia, ciência da computação e biologia. Nesta dissertação, o estudo da confiança dos dados será realizado especificamente no âmbito das aplicações de *Business Intelligence* e seu componente principal, i.e., *Data Warehouse*.

Embora exista uma crescente literatura sobre teorias e aplicações de sistemas de confiança e de reputação, definições nem sempre são coerentes entre si [5, 46]. No entanto, o conceito da confiança é, sem dúvida, associado com o conceito de confiabilidade [47], conforme defende Rodriguez et al. [6] em seu trabalho a respeito da incerteza da qualidade, inerente aos ambientes de BI:

- A confiança é a probabilidade subjetiva pela qual um partido espera que outro partido realize uma determinada ação em que seu bem-estar ou de negócios depende [46, 48];
- A reputação é a opinião geral sobre uma pessoa, empresa ou objeto. Assim, enquanto a confiança deriva de fenômenos pessoais e subjetivos, podemos considerar a reputação uma medida coletiva de confiabilidade, baseada nas referências ou nas classificações de membros de uma comunidade.

Para os cientistas da computação, confiança e reputação são particularmente importantes para apoiar decisões no fornecimento de serviços baseados na Internet. Reputação, especialmente, pode conduzir a relações dos indivíduos e das empresas em mercados *online* [49,50]. Por exemplo, eles podem usar sistemas de filtragem colaborativa para julgar o comportamento de um partido e ajudar outros partidos na decisão de iniciar um negócio com essa parte. Um sistema de reputação coleta, agrega e distribui *feedbacks* sobre o comportamento dos participantes do passado e desencoraja o comportamento injusto [51]. A análise cruzada de diferentes sistemas de reputação nos permite perceber os mecanismos e métodos para a monitoração e melhoria da reputação *online* [6, 52].

3.3.1. Abordagens para representação da confiança

Nosso objetivo nesta seção é estabelecer uma classificação das abordagens relacionadas com a questão da confiabilidade na literatura. No que se refere à qualidade de dados, a noção da confiabilidade pode ser classificada em três grupos: (a) como um atributo (ou dimensão) da QD, refletindo a reputação das fontes de dados [7, 8, 28]; (b) confiança como uma categoria composta de várias dimensões, representando um agregado de múltiplos fatores de confiança nos dados, e.g., acurácia e linhagem dos dados [25, 26, 27] e (c) como atributo chave da incerteza dos dados [9].

3.3.2. Fonte dos dados: A visão tradicional da confiabilidade na QD

A visão tradicional do atributo confiabilidade encontra-se relacionada diretamente a com a reputação das fontes de dados. Embora todas tenham o aspecto do agrupamento de dimensões de QD em comum, muitas divergem entre si por agrupar os atributos de diferentes formas.

Uma das abordagens mais citadas na literatura é a de Wang e Strong [8] que, em 1996, conduziram a primeira pesquisa empírica em larga escala sobre a QD. Seus objetivos eram identificar as dimensões de qualidade de dados na percepção dos consumidores, bem como a importância de cada atributo (137 profissionais e acadêmicos foram consultados).

Após o levantamento das dimensões de qualidade, os atributos foram reunidos em quatro categorias: Intrínseca, Contextual, Operacional e Representacional. O resultado deste trabalho encontra-se na Tabela 3.2, através de uma descrição das categorias, e na Tabela 3.3, no agrupamento das dimensões de QD [8, 29]. É possível observar que nesta abordagem, a confiabilidade encontra-se representada pela dimensão "reputação", na categoria "Intrínseca", e pela "confiabilidade" na categoria "Contextual". Está representando, portanto, a expectativa do bom funcionamento da arquitetura de *hardware* e de *software* do Sistema de Informação.

Tabela 3.2: Categorias da QD (ponto de vista do usuário)

Categorias de QD	**Descrição**
Intrínseca	Denota que os dados tenham qualidade, por direito próprio (uma característica de sua natureza)
Operacional	Enfatiza a importância do papel do sistema, refletindo a necessidade do sistema ser acessível e ao mesmo tempo seguro
Contextual	Ressalta que a qualidade dos dados deve ser considerada dentro do contexto do trabalho a ser realizado
Representacional	Denota que o sistema deve apresentar os dados de uma forma interpretável

Tabela 3.3: Categorias e dimensões da QD (ponto de vista do usuário)

Categorias de QD	Dimensões
Intrínseca	Acurácia, objetividade, credibilidade, reputação, atualidade, duplicidade, validade da rastreabilidade
Operacional	Acessibilidade, segurança, interatividade, disponibilidade, suporte ao cliente, facilidade de operação, tempo de resposta
Contextual	Aplicabilidade, completude, flexibilidade, novidade, confiabilidade, relevância, especialização, atualizada, com valor apropriado, com valor agregado
Representacional	Interpretável, com facilidade de entendimento, representação concisa, com representação consistente, em quantidade de dados suficiente, atratividade, documentação, organização

3.3.3. Confiabilidade como um agregado de dimensões

Nesta perspectiva a importância da confiabilidade é elevada, passando a ser considerada um agregado de dimensões de QD. Diversas definições conflitantes encontram-se classificadas nesta perspectiva. Uma definição mais recente sobre a confiança considera que a qualidade de dados, por si só, não é o suficiente para atestar que os dados encontram-se em um nível suficientemente satisfatório para o usuário. Aliado a esta questão encontra-se a necessidade emergente por soluções holísticas para a avaliação da QD [1]. Por meio de uma revisão na literatura, Gamble e Goble (2011) reagruparam os atributos de qualidade de dados em três categorias, e os ordenaram pelo número de citações em estudos de Qualidade de Dados e Qualidade da Informação (QI) para avaliação de dados científicos da *web*: qualidade, confiança e utilidade [27], conforme Tabela 3.4:

Tabela 3.4: Análise das dimensões da QI. Adaptado de (Gamble e Goble, 2011)

Dimensões da Qualidade	Indicador de	Função do	No. de citações em estudos de QD/QI
Completude	Qualidade	Artefato/Padrão	9
Acurácia	Qualidade	Artefato/Processo/ Padrão	9
Atualidade	Utilidade	Artefato/Consumidor	8
Consistência	Qualidade	Artefato/Processo/Padrão	8
Acessibilidade/Disponibilidade	Utilidade	Artefato/Consumidor	7
Reputação	Confiança	Artefato/Produtor/Provedor/Processo /Consumidor	7

(continua)

Dimensões da Qualidade *(continuação)*	Indicador de	Função do	No. de citações em estudos de QD/QI
Objetividade	Confiança	Artefato/Produtor/Provedor/ Consumidor	7
Conciso	Utilidade	Artefato/Consumidor	6
Relevância	Utilidade	Artefato/Consumidor	6
Compreensibilidade	Utilidade	Artefato/Consumidor	6
Credibilidade	Confiança	Artefato/Produtor/Provedor/ Consumidor	5
Interpretabilidade	Utilidade	Artefato/Consumidor	5
Dado corrente	Qualidade	Artefato/ Padrão	5
Segurança	Confiança	Artefato/Produtor/Provedor/ Processo/ Consumidor	4
Quantidade de dados	Utilidade	Artefato/Consumidor	4
Exatidão	Qualidade	Artefato/ Padrão	4
Valor Agregado	Utilidade	Artefato/Consumidor	3
Estabilidade / Volatilidade	Qualidade	Artefato/Processo/ Padrão	3
Aplicabilidade / Adequação	Utilidade	Artefato/Consumidor	2
Autoridade	Confiança	Produtor/Provedor	2
Livre de Erros	Qualidade	Artefato/ Padrão	2
Recomendação	Confiança	Artefato/Produtor/Provedor/ Processo	2
Confiabilidade	Confiança	Produtor/Provedor/Consumidor	2
Utilidade	Utilidade	Artefato/Consumidor	2
Custo	Utilidade	Artefato/Consumidor	2
Usabilidade	Utilidade	Artefato/Consumidor	2

Na visão de Gamble e Goble [27], a confiança, assim como as demais dimensões da QD, são uma função de seis entidades únicas com potenciais de avaliação: (i) artefato (pode ser entendendido, em nosso contexto de aplicação, como o dado em si), (ii) produtor, (iii) provedor, (iv) processo, (v) consumidor dos dados e (vi) um padrão referência de qualidade. Esta conceituação apresenta-se tendenciosa ao contexto de dados científicos na *web*. O cálculo geral passa pela qualidade, confiança e utilidade através de uma rede de decisão, conforme Figura 3.1.

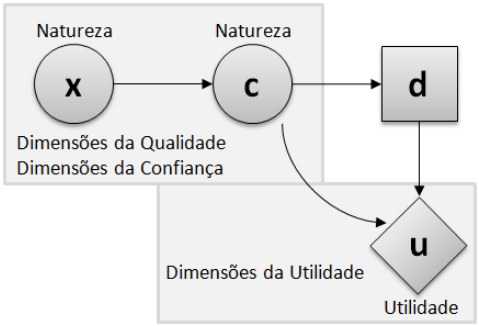


Figura 3.1: Rede de Decisão para avaliação da QD da *web*. Adaptado de (Gamble e Goble, 2011)

A investigação da confiabilidade através da linhagem (ou proveniência) dos dados foi abordada em [73]. Nessa investigação, o processo de avaliação da confiabilidade dos dados deu-se através da observação do dado a partir de sua origem, eventuais transformações entre intermediários e sua disponibilização ao consumidor final. Nesta investigação, os dados e os fornecedores são passíveis de medição de confiabilidade. Esta proposta, juntamente com a [6] demonstraram-se particularmente úteis em repositórios de dados no qual a entrada dos dados é realizada sem qualquer controle de qualidade (e.g., prontuários hospitalares, formulários inscrição manuais, observações de agentes de *software*).

Utilizando outro juízo crítico, Pawluk [25] elevou mais ainda o conceito da confiança em sua abordagem de para Gerenciamento de Dados Mestres *(Master Data Management - MDM)*. Em seu trabalho, ele argumenta que a confiança de dados estende o conceito da qualidade de dados por meio de diversas dimensões. A conceituação parte da definição de QD de Naumann [39], como uma tentativa de fornecer uma definição operacional da QD através de um valor agregado de múltiplos critérios de Qualidade da Informação. Algumas definições sobre a confiança de dados são apresentadas pelo próprio Pawluk [26].

Definição 3.3.3.1. *(Confiança dos Dados): A confiança é o valor agregado dos múltiplos fatores de Confiança dos Dados.*

O autor indica, não explicitamente, que uma característica mais geral a respeito dos dados é composta de $confiança + qualidade$. A Definição 3.3.3.1 mencionada acima proporciona flexibilidade na definição de confiança para um setor específico e para as necessidades dos utilizadores. O fator de confiança dos dados ($DT - factor$) pode ser um fator de QD ($QD - factor$), ou não fator de qualidade (NQ). Em seguida, o autor conclui

apresentando a acurácia como atributo essencial para medição da confiança nos dados, conforme Definição 3.3.3.2.

Definição 3.3.3.2. *(Medição da Acurácia): É considerada a medida acurada se e somente se a chave da tupla* $(x.X_0)$ *for precisa e o valor da medida* $(x.X_i)$ *for igual ao valor do mundo real* $(x.X_i)$ *identificado pela chave da tupla que pertence tanto a chave quanto a medição da Equação 3.1:*

$$Acc(x.X_i) = \begin{cases} 1 & if \quad x.X_i = x'.X_i \wedge Acc(x.X_0) = 1 \\ 0 & if \quad x.X_i \neq x'.X_i \vee Acc(x.X_0) = 0 \end{cases} \qquad \text{(Eq. 3.1)}$$

Outras pesquisas relevantes são fornecidas pela comunidade de QD onde a confiabilidade dos dados é muitas vezes considerada sinônimo de credibilidade [39]. Em [76] a decomposição da credibilidade é apresentada em três sub-dimensões: confiabilidade da fonte, razoabilidade dos dados e a temporalidade dos dados. Seguindo essa diferenciação, Prat e Madnick [74] propõem uma abordagem baseada na proveniência para medir credibilidade, agregando escores de qualidade para as subdimensões.

3.3.4. A confiança como atributo chave para o cálculo da Incerteza dos Dados

A proposta do cálculo da incerteza dos dados é apresentada por Rodriguez et al. [6]. Nesta perspectiva, a questão da incerteza está relacionada com a confiança, no que se refere aos processos de negócio. Processos com total visibilidade (internos, no contexto corporativo) podem ser considerados confiáveis. Processos externos, nos quais não existe a visibilidade, são pouco confiáveis. A pesquisa não se concentrou em como calcular incertezas para eventos individuais. Em vez disso, priorizou o problema de como calcular e representar a incerteza ao exibir os dados.

Eles defendem que, mesmo em cenários fechados (i.e., com total visibilidade), muitas fontes possíveis de incerteza permanecem em aplicações de BI. Esse problema é ampliado quando os dados provêm de várias fontes, é coletado com diferentes métodos e freqüências e por diferentes departamentos, instituições e geografias. Em alguns casos, podemos facilmente prever ou detectar a incerteza (e.g., um parceiro não envia os dados dentro do prazo, ou uma fonte tem um método de coleta de dados inerentemente não confiável), enquanto que em outros, os problemas são pontuais e mais difíceis de reconhecer. Neste trabalho, três situações, na qual a incerteza encontra-se presente foram

identificadas: (a) registros errados no sistema, (b) base de dados incompleta e (c) dados inconsistentes, por conta de duplicidade.

Para lidar com o desafio, a noção da incerteza baseada em eventos de *logs* de dados e é composta de três atributos: (a) confiança, (b) completude e (c) acurácia. Por fim, é proposta a visualização dos dados na interface de BI, juntamente com o cálculo da incerteza. Neste modelo, não são apresentados métodos e técnicas consistentes para o calculo da incerteza.

3.4. TAXONOMIA DA CONFIANÇA DOS DADOS

Uma abordagem possível para gerar uma medida conjunta, ou métricas para refletir essa inter-relação entre confiança e qualidade dos dados foi inicialmente discutida em [4]. Neste trabalho, os valores de confiança refletem a qualidade e os valores de qualidade refletem a confiança. Em [3] é proposto definições pragmáticas e praticáveis (portanto, não apenas filosóficas) para a noção da confiança e de outros conceitos que esta depende. A proposta de uma taxonomia/fluxo de definições sobre a confiança é ilustrada na Figura 3.2 por Kashyap (2004).

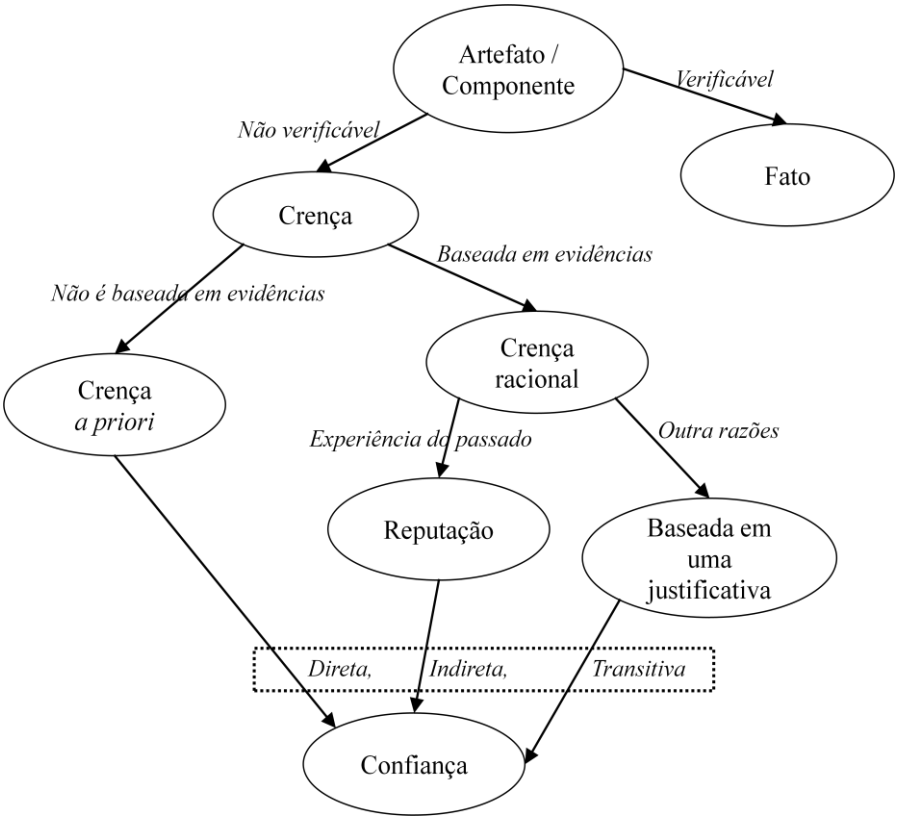


Figura 3.2: A taxonomia do fato, crença e confiança. Adaptado de (Kashyap, 2004)

A descrição de cada uma das entidades é apresentada a seguir [3]:

- **Fato:** Se o dado/informação pode ser verificado por algum meio, então ele é definido como um "fato". A verificabilidade pode ser realizada por teorias (e.g., $2 + 2 = 4$), ou com base na "confiança" em uma autoridade (e.g., uma afirmação de algum *site* do Governo). Caso isto não seja possível, temos que nos basear na "crença" dos fatos.

- **Crença:** Ocorre quando assumimos a validade de algum pedaço de dado/informação ou a qualidade de um serviço sem a verificação objetiva, ou sem recorrer a uma autoridade confiável. De acordo com o dicionário *Webster* (EUA), o termo crença pode ser entendido como um estado ou hábito de espírito em que a confiança é colocada em alguma pessoa ou coisa. É, portanto, uma noção intimamente ligada à confiança.

- **Crença *A priori*:** É caracterizada quando assumimos a validade de alguns dados/informações de uma forma *a priori*, sem uma lógica subjacente (na fronteira com a fé). Essa noção parece bíblica, mas uma análise de como os sistemas inteligentes são usados ou implantados pode revelar que essa noção é utilizada com bastante frequência.

- **Crença racional:** Quando a validade de alguns dados/informações é assumida com base em alguma lógica ou justificação, este é indicado como crença racional. Um exemplo é o da **reputação** de um sistema inteligente, o qual é baseado em experiências passadas com o sistema. Essas experiências podem ser medidas de forma objetiva e empírica. Pode-se recorrer, por exemplo, a medidas como a consistência e a completude, ou ainda a qualidade do serviço. Estas dimensões são tipicamente utilizadas na literatura da qualidade dos dados para caracterizar os resultados passados, recebidos a partir de um sistema de informação.

- **Confiança:** A noção de confiança se baseia na continuidade de noções de crença que são definidos acima.

Assim, pode-se ter a confiança *a priori* em algumas informações/dados ou serviços. Essa confiança *a priori* poderia ser muito útil nos estágios iniciais (do termo inglês *bootstrapping)* e poderia ser modulada com medidas mais objetivas, tais como a

consistência, a completude e a qualidade do serviço. Finalmente, a confiança poderia ser direta (eu confio nesta informação) ou indireta/transitiva (confio nesta informação porque meus amigos confiam nela) [14].

Ainda segundo Kashyap [3], nesse contexto, a métrica da confiança/qualidade poderia ser vista como uma medida multidimensional. Um conjunto de dimensões poderia descrever medidas objetivas (ou verificáveis), e outra poderia compreender de medidas subjetivas (ou *a priori*). Estas dimensões podem ser combinadas de duas maneiras:

- **Comparação Multidimensional:** As dimensões podem ser priorizadas, e.g., de acordo com uma prioridade mais alta para as dimensões objetivas. Assim, uma função de comparação pode escolher comparar primeiro as dimensões objetivas. Caso as dimensões objetivas sejam iguais ou dentro de uma margem de erro estatístico, as dimensões subjetivas poderiam ser comparadas.
- **Composição em uma única medida:** A outra abordagem é combinar os valores de dimensões em uma única medida comum, através da atribuição de pesos para as várias dimensões. Por exemplo, pode-se combinar uma medida subjetiva, como Crença *a priori*, e uma medida objetiva, como a Reputação, da forma apresentada na Equação 3.2:

$$Qualidade = \alpha * CrençaApriori + \beta * Reputação \quad \text{(Eq. 3.2)}$$

Os pesos podem refletir requisitos da aplicação. Por exemplo, na ausência de qualquer outra informação, pode-se optar por definir $\alpha = 1, \beta = 0$. Além disso, a Crença *a priori* pode ter em si vários níveis de subjetividade e a Reputação pode ser caracterizável em termos de fatores de dado/informação de qualidade, tais como consistência e completude; ou qualidade dos fatores de serviço, conforme Equação abaixo:

$$Qualidade = \alpha_1 * T_d + \alpha_2 * T_{id} + \beta_1 * Cons + \beta_2 * Comp + \beta_3 * Q_{Serv}, \quad \text{(Eq. 3.3)}$$

Onde,

$T_d = confiança\ direta$
$T_{id} = confiança\ indireta$
$Cons = consistência$
$Comp = completude$
$Q_{Serv} = qualidade\ do\ serviço$
$\alpha, \beta = fatores\ de\ ajuste$

Entretanto, conforme o autor [3] adverte, mais experimentos seriam necessários para aperfeiçoar essas equações empíricas e compreender a verdadeira natureza da inter-relação entre confiança e qualidade.

3.5. MÉTRICA GLOBAL DE CONFIANÇA DOS DADOS

Um número crescente de dimensões de QD tem sido identificado nos últimos anos, tornando o gerenciamento cada vez mais complexo devido às relações de interdependência, ambigüidades e duplicidades.

Conseqüentemente, vários trabalhos têm investigados métodos que combinam estes critérios e, portanto, simplificam a visão multidimensional da qualidade. Diversas técnicas de agregação têm sido sugeridas e exploradas de modo a obter uma pontuação global (ou métrica global) para a avaliação da qualidade. Podemos classificá-las em dois grupos principais, nos métodos aditivos e não aditivos [69]:

- **Funções de agregação aditiva:** Que se dedicam a resumir as medidas comensuráveis através de uma função contínua crescente, tais como média aritmética simples e ponderada [28];
- **Funções de agregação não aditiva:** Que buscam uma declaração representante do conjunto de critérios subjacente por computação, ou uma Função de Crença ou uma Função de Utilidade, tais como a Média Ponderada Ordenada *(Ordered Weighted Average - OWA),* Funções Máximo ou Mínimo, e mais recentemente, técnicas como a integral de Choquet [69], booleano, teorema de Bayes [30], redes neurais artificiais, entre outros [28, 31].

Apesar de sua simplicidade, funções aditivas implicam em restrições à natureza das medidas agregadas. Elas, na verdade, supõem que as medidas são independentes. Ou seja, não são influenciáveis por fenômenos nem conflitos, nem consideram qualquer sinergia entre os indicadores.

Para resolver questões do mundo real que necessitem de estruturas matemáticas e computacionais capazes de lidar com imprecisões e incertezas, de forma mais crítica e realista, Zadeh formalizou, na década de 60, um conceito revolucionário que sofreu, naturalmente, muitas resistências. Entretanto, com o tempo, sua proposta mostrou-se prática e capaz de auxiliar na modelagem de situações nas mais variadas áreas do

conhecimento. Esse conceito foi chamado de Teoria dos Conjuntos Difusos ou Lógica *Fuzzy* [70].

3.5.1. A incerteza inerente a Confiabilidade dos Dados

Existem domínios de aplicação nos quais a incerteza é parte inerente do problema, devido a dados ausentes ou imprecisos e/ou relações causa-efeito não determinísticas. Ambientes de raciocínio com incerteza exigem [75]:

- Quantificação de Incerteza;
- Método de combinação dos valores de Incerteza.

Na literatura, a imperfeição da informação é geralmente conhecida como imprecisão ou incerteza. Estes dois elementos permeiam os cenários do mundo real e devem ser incorporados em todos os sistemas de informação que tentam oferecer um modelo completo e acurado do mundo real. Dados, como os disponíveis em um sistema de informação, são sempre de alguma forma imperfeitos [77].

Para tratar a incerteza, é necessário encontrar um modelo adequado para representar a informação imperfeita, de acordo com o seu tipo de imperfeição:

- Informação probabilistica: Teoria de probabilidades ou teoria da evidência;
- Informação imprecisa ou vaga: Teoria de conjuntos *fuzzy* ou teoria de conjuntos de aproximação *(rough set)*;
- Informação possibilista: Teoria de possibilidades;
- Informação incerta: Teoria de probabidades, teoria de possibilidaes ou teoria da evidência.

Para suportar o raciocínio impreciso e incerto em relação à informação disponibilizada ao usuário, nosso modelo de confiabilidade será baseado na lógica *fuzzy.*

3.5.2. Lógica Fuzzy

Aristóteles foi o fundador da ciência lógica, criando a lógica Aristotélica ou lógica bivalente. O emprego da lógica Aristotélica levava a uma linha de raciocínio lógico baseado em premissas e conclusões, na qual determinada informação só poderia ser

verdadeira (1) ou não verdadeira (0). Porém Bartrand Russell mostrou através do "Paradoxo de Russell" que nem todos os problemas poderiam ser resolvidos pela lógica bivalente. Em torno de 1930, Jan Lukasiewicz desenvolveu a lógica multinível, onde uma determinada afirmação poderia ser verdadeira e falsa ao mesmo tempo. Isso se torna possível desde que não apresente apenas dois níveis, verdadeiro e falso, mas sim um grau de verdade, existindo, portanto, vários níveis [75].

A teoria dos conjuntos nebulosos (ou conjuntos *fuzzy*) foi desenvolvida a partir de 1965, por Lotfi A. Zadeh (professor da Universidade da Califórnia, Berkeley - EUA), baseado na lógica multinível [70]. O objetivo deste trabalho era fornecer um ferramental matemático para o tratamento de informações de caráter impreciso ou vago. Foi a partir do mesmo que surgiu a expressão lógica *fuzzy*, onde o termo em inglês *"fuzzy"*, traduzido para o português, tem o significado de algo nebuloso ou difuso.

Hoje em dia ela encontra-se entre as técnicas mais populares de Inteligência Artificial e vêm sendo amplamente aplicada com sucesso em diversas áreas, como: automação e controle, classificação e reconhecimento de padrões, tomada de decisão, sistemas inteligentes, previsão de séries temporais, robótica, entre outras [75, 82].

Comercialmente, lógica *fuzzy* tem sido explorada em diversos produtos, e.g., metrô da cidade de Sendai (Japão) e de São Paulo (Brasil), lavadoras de roupa, máquinas filmadoras, aspiradores de pó, fornos de microondas, ar condicionado, freios ABS (acrônimo para a expressão alemã *Antiblockier-Bremssystem*), entre outros.

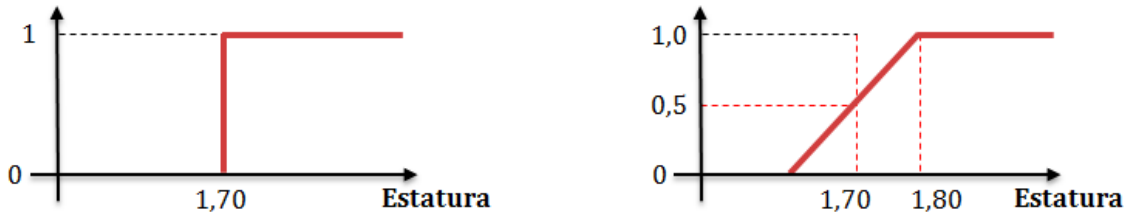
Formalmente, um conjunto nebuloso A do universo de discurso Ω é definido por uma função de pertinência $\mu_A : \Omega \rightarrow [0,1]$. Essa função associa a cada elemento x de Ω o grau $\mu_A(x)$, com o qual x pertence a A [1]. A função de pertinência $\mu_A(x)$ indica o grau de compatibilidade entre x e o conceito expresso por A:

- $\mu_A(x) = 1$ indica que x é completamente compatível com A;
- $\mu_A(x) = 0$ indica que x é completamente incompatível com A;
- $0 < \mu_A(x) < 1$ indica que x é parcialmente compatível com A, com grau $\mu_A(x)$.

Um conjunto A da teoria dos conjuntos clássica pode ser vista como um conjunto nebuloso específico, denominado usualmente de conjunto *"crisp"* (termo inglês, que pode

ser traduzido como clássico), para o qual $\mu_A : \Omega \rightarrow \{0,1\}$, ou seja, a pertinência é toda do tipo "sim ou não", "certo ou errado", e não gradual, como para os conjuntos nebulosos.

A diferença entre estes dois conceitos em relação a variável "estatura" é ilustrada na Figura 3.3a e na Figura 3.3b, que descrevem, respectivamente, a representação do conceito "alto" através de um conjunto clássico (ou *crisp*) e de um conjunto nebuloso (ou *fuzzy*).



(a) **(b)**

Figura 3.3: Conjuntos clássico e nebuloso para a variável "estatura" e seu valor lingüístico "alto". (a) Conjunto Clássico. (b) Conjunto Nebuloso.

O conjunto clássico *A* da Figura 3.3a, não exprime completamente o conceito de "alto", pois uma pessoa com estatura 1,69 seria considerada completamente incompatível com este conceito. Na verdade, qualquer intervalo clássico que se tome para representar este conceito é arbitrário.

Já o conjunto nebuloso *B* da Figura 3.3b permite exprimir que qualquer pessoa com estatura a partir de 1,80 seja considerada "alta"; abaixo de 1,60 não seja considerado "alto", e no intervalo {1.60, 1.80} seja considerada tanto mais "alto" quanto mais próximo de 1,80 é sua estatura.

A cardinalidade de um conjunto nebuloso *A* é expressa pelas Equações 3.4 e 3.5:

- Para Ω discreto:

$$|A| = \sum_{x \in \Omega} \mu_A(x) \qquad \text{(Eq. 3.4)}$$

- Para Ω contínuo:

$$|A| = \int_{\Omega} \mu_A(x) \qquad \text{(Eq. 3.5)}$$

Os formatos mais usuais desta função são e.g., a função triangular, a trapezoidal e a Gaussiana. De acordo com [75], um sistema de inferência *fuzzy* é composto de quatro componentes, obedecendo ao diagrama apresentado na Figura 3.4.

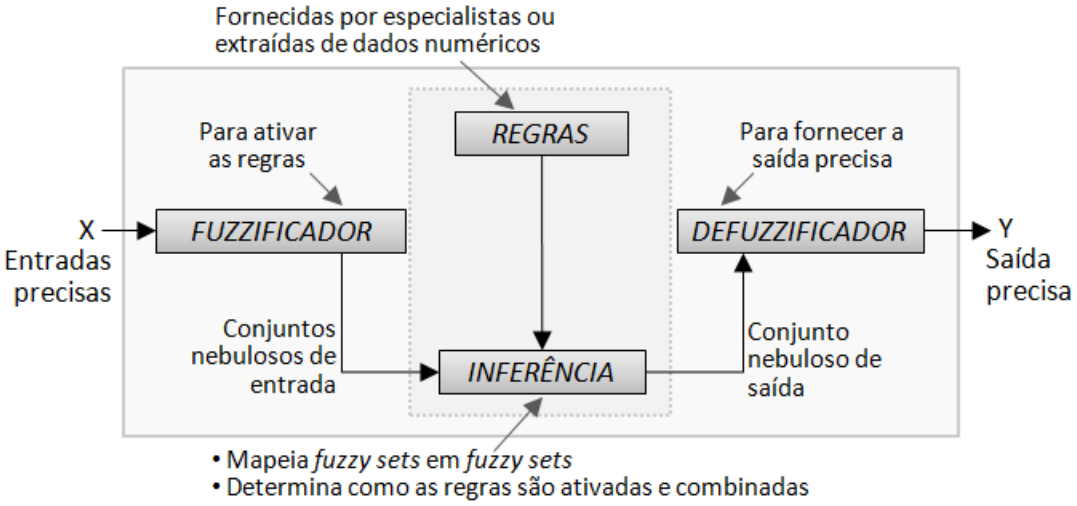


Figura 3.4: Arquitetura de um sistema de inferência fuzzy

- **Fuzzificador**: É o processo de associar ou calcular um valor para representar um grau de pertinência dos valores reais de entrada, em um ou mais grupos qualitativos, chamados de conjuntos *fuzzy*. O grau de pertinência é determinado por uma função de pertinência. Dentre os *fuzzificadores* mais utilizados [75], destacam-se *singleton* (ou singular), Gaussiano, Triangular e Trapezoidal.

- **Defuzzificador**: É definido como uma função que associa a cada conjunto *fuzzy* um valor real. O valor escolhido pode ser entendido como uma espécie de valor esperado, traçando uma analogia com as distribuições de probabilidade. Existem vários métodos de *deffuzificação* (pelos menos 30 tipos diferentes são citados na literatura, e.g., Centro Ponderado, Máximo, etc.) e o método *COG (Center of Gravity* ou método do centróide, ou ainda Centro de Gravidade) é o mais comumente adotado. Ele fornece um valor correspondente à abscissa do baricentro do gráfico da função de pertinência. A Equação 3.6 é usada para o cálculo:

$$R_{ij} = \sum_{j=1}^{k} w_j * r_{ij} \Big/ \sum_{j=1}^{k} w_j \qquad \text{(Eq. 3.6)}$$

Onde,

w_j são os pesos *fuzzy* dos atributos;

r_{ij} é o grau de atendimento de cada atributo da característica avaliada; e

R_{ij} é grau de atendimento do ambiente a um padrão determinado.

- **Base de Regras Fuzzy:** Armazena o conhecimento humano, que consiste de uma base de regras, de maneira a caracterizar a estratégia de controle. Na base de regras ficam armazenadas as definições sobre a discretização e a normalização dos universos de discurso, bem como as definições das funções de pertinência dos termos nebulosos. A base de regras é formada por estruturas do tipo:

  Se <*premissa*> **Então** <*conclusão*> *("IF-THEN")*

  Sendo que, essas regras, juntamente com os dados de entrada, são processadas pelo procedimento de inferência, o qual infere as ações de controle de acordo com o estado do sistema.

- **Máquina de Inferência Fuzzy**: É responsável por combinar as regras *fuzzy "IF-THEN"* existentes na base de regras em um mapeamento de um conjunto *fuzzy* de saída. Os tipos de controladores *fuzzy* encontrados na literatura são os modelos clássicos, compreendendo o modelo de Mamdani e o de Larsen, e os modelos de interpolação, compreendendo o modelo de Takagi-Sugeno e o de Tsukamoto [75]. Os modelos diferem quanto à forma de representação dos termos nas premissas, quanto à representação das ações de controle e no que se refere aos operadores utilizados para a implementação do controlador.

Similarmente às diversas áreas da Inteligência Artificial, a Lógica bem abordou a definição de sistemas de inferência que levam em conta elementos de situações da vida real. Medidas que objetivam a incerteza formalizam a força de nossas crenças na ocorrência de alguns eventos atribuindo, a estes eventos, um grau de crença sobre a sua ocorrência [78], i.e., problemas de qualidade de dados em nosso contexto.

Baseado nas contribuições expostas até aqui, e em outros modelos encontrados na literatura, a proposta para avaliação da confiabilidade dos dados, no âmbito das aplicações de BI, é apresentada no Capítulo 4.

4. MODELO FUZZY DE CONFIABILIDADE DE DADOS PARA AMBIENTES DE BI

Neste capítulo serão apresentadas as características do modelo *fuzzy* para a avaliação da confiabilidade dos dados em ambientes de BI, baseada em uma taxonomia dos problemas de QD em DW. Em razão do que foi destacado nos capítulos anteriores, a confiabilidade dos dados em ambientes de BI é, portanto, um importante aspecto a ser investigado. A solução holística proposta para a QD está organizada em quatro etapas, conforme ilustrado na Figura 4.1.

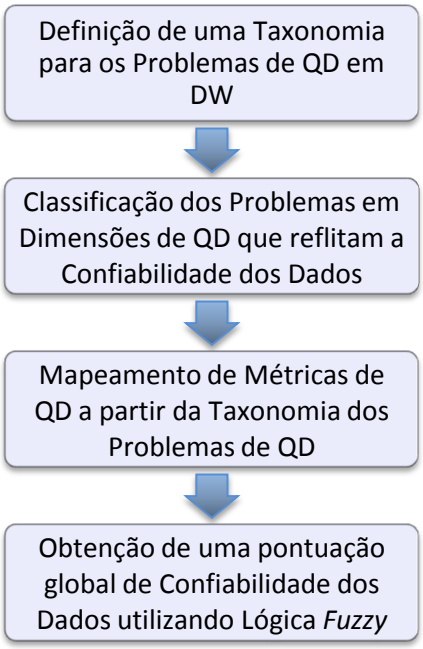


Figura 4.1: Processo de Desenvolvimento da Solução

- **Definição de uma Taxonomia para os Problemas de QD em DW:** Desenvolvimento de uma nova taxonomia, com o objetivo de descrever os problemas de QD passíveis de observação em ambientes de DW/BI, levando em conta os diferentes níveis de granularidade no qual o problema se manifesta (i.e., valor individual, multi-valor, tupla, coluna e relação Fato-Dimensão).
- **Classificação dos Problemas em Dimensões de QD que reflitam a Confiabilidade dos Dados:** Agrupamento dos Problemas de QD identificados, com base em suas características comuns, em dimensões que reflitam os aspectos da confiabilidade dos dados em ambientes de DW/BI;

- **Mapeamento de Métricas de QD a partir da Taxonomia dos Problemas de QD:** Definição de métricas que mensurem o grau de presença de determinado problema de QD nos DW/BI. Como veremos a seguir, a incerteza inerente a qualidade dos dados encontra-se na dificuldade (e, em alguns casos, na impossibilidade) de mapeamento de determinados problemas de QD em métricas, fatos e/ou métodos de medição;

- **Obtenção de uma pontuação global de Confiabilidade dos Dados utilizando Lógica *Fuzzy*:** A partir dos Problemas de QD, e de suas respectivas métricas agrupadas em diferentes dimensões, será possível obter uma pontuação global para o nível de confiabilidade dos dados.

Este capítulo está organizado da seguinte maneira. A Seção 4.1 apresenta a definição para a confiabilidade dos dados. Em seguida, a Seção 4.2 descreve uma nova taxonomia para os problemas de QD em DW. Na Seção 4.3, métricas para avaliação da qualidade dos dados são propostas a partir da investigação dos problemas de qualidade. Esta análise é, em seguida, estendida através de um método de agregação das métricas em dimensões de qualidade na Seção 4.4. Por fim, na Seção 4.5, um modelo para representar a noção da confiabilidade dos dados é proposto baseado na lógica *fuzzy*.

4.1. DEFINIÇÃO PARA A CONFIABILIDADE DOS DADOS

Os termos qualidade e confiança estão relacionados, mas não refletem necessariamente a mesma coisa. Esta investigação parte da idéia de [26], no qual baixa qualidade dos dados pode ser confiável em algumas situações e dados de alta qualidade podem ter baixa confiança em outro contexto.

Para ilustrar este pensamento, considere, por exemplo, a ocorrência de dados duplicados. Tais situações geram inconsistência e, conseqüentemente, diminuem a qualidade dos dados. Entretanto, sob a perspectiva da confiança, é possível afirmar que a informação é mais confiável do que as demais que não apresentam uma violação na restrição de unicidade. É, portanto, neste exemplo, como se duas pessoas afirmassem um fato, ao invés de apenas uma.

Ao examinar a confiabilidade de uma informação fornecida, é indispensável a preocupação com os dados utilizados, em virtude dos diversos problemas de qualidade

existentes. Neste contexto, dois conceitos importantes e inter-relacionados em ambientes de BI (a exemplo de trabalhos em outros domínios, e.g., em sistemas *Peer-to-Peer (P2P)* [76]) são a confiança e a confiabilidade.

A confiança é definida como a crença do usuário na qualidade dos dados na fonte e nas transformações sofridas durante o processamento, com o propósito de fornecer informações como o esperado. Para caracterizar a confiança, temos de ter alguma forma de medi-la. A confiabilidade pode ser entendida como uma medida que mostra o nível de confiança que o usuário tem sobre uma informação fornecida pelo ambiente de BI, num dado contexto, em um intervalo de tempo determinado.

A seguir, apresenta-se a definição formal para a confiabilidade dos dados:

Definição 4.1.1. *A Confiabilidade de Dados é um valor agregado de múltiplos fatores de Confiabilidade, denotado como* $conf(v_i)$*, e é a probabilidade de* v_i *estar correto, de acordo com o melhor de nosso conhecimento.*

A Definição 4.1.1 acima mencionada realça a necessidade de abordagens suficientemente capazes de balizar uma impressão bem fundamentada acerca dos dados armazenados em ambientes de DW/BI. Diferente das metodologias tradicionais, preocupadas em medir a qualidade/confiança a partir do mapeamento dos fatores de qualidade (ou seja, em um alto nível de abstração) e freqüentemente construídos a partir de uma base *ad hoc* para resolver problemas específicos e práticos, este trabalho parte de uma verificação exaustiva e sistemática dos problemas de QD manifestos em DW/BI. O resultado de nossa investigação é apresentado na seção seguinte, sob a estrutura de uma Taxonomia dos Problemas de QD em DW.

4.2. TAXONOMIA DOS PROBLEMAS DE QD EM DW

A primeira contribuição deste trabalho é apresentada nesta seção. O desenvolvimento de uma taxonomia para os problemas de qualidade de dados em DW/BI é proposto com o objetivo suprir uma lacuna de modelos que facilitem a compreensão, gestão e avaliação dos problemas de qualidade. Seguidamente, a taxonomia é comparada com os trabalhos relacionados, evidenciando ser mais abrangente e indicada para ambientes de DW que as demais existentes na literatura para a cobertura dos problemas de QD.

Um padrão característico das investigações de medição da qualidade reside em sua abordagem tipicamente *top-down*, conforme ilustra a Figura 4.2. Nela, é possível observar que a qualidade dos dados é composta de quatro níveis de abstração (Etcheverry et al., 2008) [72].

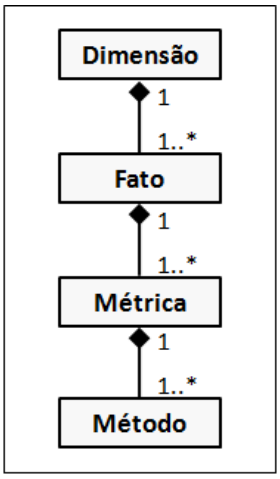


Figura 4.2: Níveis de abstração da QD. Adaptado de (Etcheverry et al., 2008)

Diferentemente dos trabalhos atuais, a metodologia apresentada neste trabalho é uma solução holística que se assenta em uma aproximação seqüencial ascendente (i.e., *bottom-up*), a fim de identificar *a posteriori* um conjunto de dimensões compostas de métricas capazes de refletir a confiabilidade dos dados. Sua concepção encontra-se, portanto, em conformidade com os níveis de granularidade que compõem a QD, partindo do nível mais elementar (i.e., extração de métricas a partir de problemas de QD) até o de maior complexidade (i.e., pontuação global da confiabilidade dos dados obtida por agregação de dimensões de QD). Não é de nosso conhecimento a existência de trabalhos que tratam da definição de métricas de qualidade a partir dos problemas recorrentes de dados sujos e/ou problemas de QD em ambientes de DW/BI. Os problemas detectados serão apresentados sob o formato de uma taxonomia, conforme considerações da seção 3.1.2.

Uma Taxonomia possui três características fundamentais: cumulatividade, hierarquia e eixo comum. Acreditamos que uma taxonomia dos problemas de QD pode ser a melhor referência para obtenção de métricas de QD pelas seguintes razões:

- Apoiar-se em um método que tem como característica inerente a definição e classificação exaustiva de todos os elementos de um determinado domínio;

- Diminuição de ocorrências de conceitos dúbios, mal-entendidos ao longo do processo de avaliação da QD em DW;
- Classificação dos problemas de QD, de forma inequívoca, em relação às dimensões de qualidade;
- Nem sempre ambientes de DW dispõem de metadados de qualidade que reflitam a linhagem dos dados, i.e., dificuldade de avaliar a qualidade em repositórios de DW já implementados;
- Apresentar-se como um método promissor para aferições de DQ, visto que o uso de uma Taxonomia de problemas de QD não tem sido explorado na literatura como ponto de partida para avaliações de qualidade.

Convém observar que este trabalho propõe uma nova perspectiva para a aferição da Confiabilidade dos Dados, baseada não nas características positivas observadas, mas sim, nos problemas de qualidade percebidos. Com base nesta afirmação, é possível considerar que a eficiência de um modelo de medição de qualidade reside em sua maior ou menor capacidade de detecção de problemas. Conforme [56] cita, poucas organizações têm dado a devida atenção para a identificação de dados sujos e a forma de tratá-los em sua base. Portanto, a partir de um mapeamento dos problemas de qualidade existentes, é possível extrair métricas correspondentes que reflitam o quão confiável os dados são.

4.2.1. Classificação dos Problemas

Diante do que foi exposto no Capítulo 3, a apresentação de uma nova taxonomia de QD em ambientes de DW/BI é apresentada a partir desta seção. Cada um dos problemas identificados foram cuidadosamente analisados e, então, agrupados em dimensões a qual o problema se identifica. Essa investigação resultou na descoberta dos múltiplos fatores de compõem a Confiabilidade de Dados, conforme Definição 4.1.

A análise foi efetuada de forma independente de qualquer ambiente de *Data Warehouse* ou *Business Intelligence* específico, o que lhe confere um caráter genérico e universal. Desta forma, a taxonomia não é meramente representativa dos problemas de QD identificados em um repositório em particular. Se assim fosse, certamente seria uma taxonomia incompleta e não representativa dos diversos problemas de QD que ocorrem.

Além do mais, esta metodologia encontra-se de acordo a taxonomia de problemas de QD de Oliveira et al. [54].

A aproximação adotada foi ascendente, i.e., começou-se por analisar o nível de granularidade mais elementar (i.e., valor individual de um atributo) e terminou-se no nível de granularidade mais complexo (i.e., relação Fato-Dimensão do modelo multidimensional). O objetivo da aproximação consistiu em identificar, em primeiro lugar, os problemas mais concretos e de fácil percepção, deixando para o fim os mais genéricos e de difícil detecção (i.e., maior subjetividade). Os problemas de QD identificados foram agrupados de acordo com suas características similares e, como resultado, encontram-se classificados e inseridos em cinco dimensões de qualidade: Completude, Atualidade, Unicidade, Consistência e Acurácia.

Nas seções seguintes, os problemas de QD identificados são apresentados, iniciando pela dimensão da Completude.

4.2.2. Problemas na Dimensão da Completude

Este grupo de problemas de QD em DW é composto por problemas que comprometem a completude dos dados no DW. O significado de cada item identificado encontra-se representado na Figura 4.3:

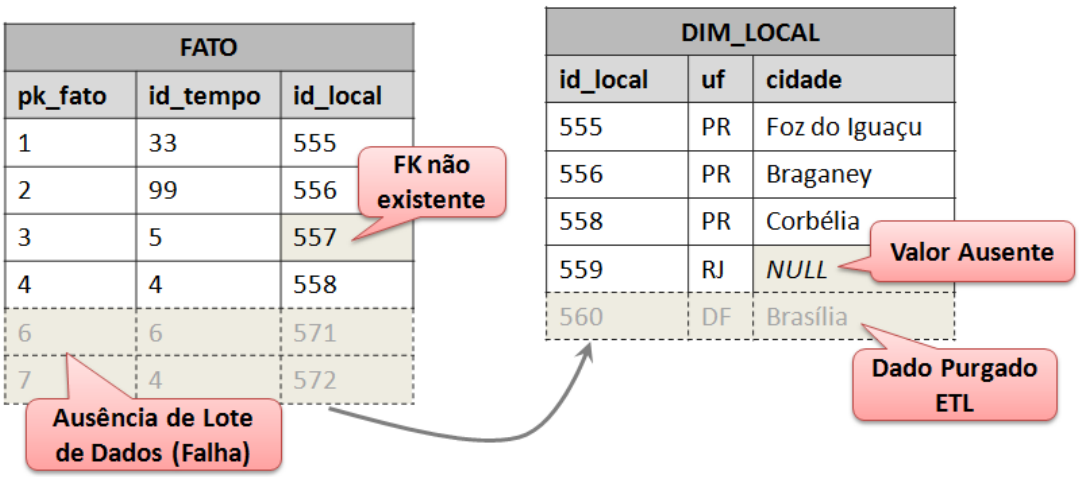


Figura 4.3: Problemas de Completude dos dados identificados

- **Valor ausente** – Ausência de valor em um atributo de preenchimento obrigatório, seja na tabela fato, ou nas dimensões (e.g., atributo com valor *NULL* ou com o valor *default*).

- **Referência definida, mas não encontrada** – Inexistência de um atributo na tabela dimensão, e que se encontra referenciada pela tabela Fato (e.g., atributo fk_Tempo da tabela fato aponta para um valor inexistente na dimensão $Tempo$).
- **Purgar dados do DW** – Dados que tiveram sua gravação no DW rejeitada durante o processo de ETL (e.g., registros eliminados durante o processo de ETL por alguma regra de qualidade).
- **Ausência de lotes de dados** – Lotes de dados ausentes no DW, devido a falhas ocorridas no processo de atualização periódica (e.g., falha no canal de comunicação entre uma fonte de dados e o repositório do DW durante a carga dos dados).

4.2.3. Problemas na Dimensão da Atualidade

Nesta seção são apresentados os problemas de QD em DW que comprometem a atualidade dos dados no DW. Os significados dos três problemas identificados são descritos a seguir:

- **Ausência de atualizações periódicas** – Ausência e/ou falhas nas atualizações periódicas automáticas do DW (e.g., ocorrência de uma falha no sistema de comunicação entre uma fonte de dados e o repositório do DW).
- **Intervalo de atualização incorreto** – Incapacidade do sistema em agendar extrações por tempo, intervalo, ou evento satisfatório as necessidades do DW (e.g., extrações agendadas para ocorrer mensalmente quando, na verdade, deveriam ocorrer semanalmente).
- **Dados desatualizados** – Dados temporais desatualizados, violando, portanto, a restrição temporal de tempo válido (e.g., um campo $cargo\_pessoa$ ou $salario\_pessoa$ não foi atualizado antes de ter sido extraído e armazenado no DW).

4.2.4. Problemas na Dimensão da Unicidade

Nesta seção são apresentados os problemas de QD em DW que comprometem a unicidade dos dados no DW. O significado de cada problema identificado é apresentado a seguir:

- **Dados duplicados** – Existência de registros e/ou tuplas duplicadas (violando a restrição de unicidade) ou com grande probabilidade (e.g. existência de duas tuplas com os mesmos valores em todos os registros na tabela Fato).
- **Dados duplicados e contraditórios** – Existência de registros e/ou tuplas duplicadas e conflitantes entre si (violando, portanto, a restrição de unicidade) ou com grande probabilidade (e.g., considere o exemplo do item anterior, na qual o repositório apresenta a existência de duas tuplas com os valores dos registros na tabela fato contraditórios entre si em determinado campo).
- **Existência de sinônimos** – Registros contendo valores diferentes, mas com significado equivalente (e.g., tabela Fato com o $cargo\_pessoa$ contendo os registros "Mestre" e "Professor").

4.2.5. Problemas na Dimensão da Consistência

Nesta seção são apresentados os problemas de QD em DW que comprometem a consistência dos dados em ambientes de DW. O significado de cada problema identificado é apresentado a seguir:

- **Modelagem do DW incompleta** – Colunas em falta nas tabelas Dimensões ou Fato do modelo físico do *Data Warehouse*, devido a uma falha por parte da equipe de definição do modelo multidimensional (e.g. Dimensão $Tempo$ do DW não contém o nível de agregação $Trimestral$ ou $Mensal$).
- **Dimensão apontada por diferentes campos** – Tabela Dimensão apontada por diferentes campos pelas Tabelas Fato, no modelo físico do repositório de dados (e.g., Dimensão $Pessoa$ apontada por ID_nome em uma tabela Fato e, pelo campo cpf_nome, por outra tabela Fato).

- **Diferentes tipos de dados para colunas semelhantes** – Existência de problemas na modelagem dos tipos de dados (e.g. chave primária é do tipo Inteiro em uma tabela, e sua chave estrangeira é do tipo *String*).
- **Diferentes representações devido a abreviações** – Violação de sintaxe, em virtude de diferentes formatos de representação dos dados não compostos, devido ao uso de abreviatura (e.g., uso de abreviações como "Dr" para **"**Doutor" ou "Diretor").
- **Diferentes representações devido a pseudômino/apelido** – Violação de sintaxe, em virtude de diferentes formatos de representação dos dados não compostos, devido ao uso de pseudômino/apelido (e.g., "Lula", "Presidente Lula", "Luiz Inácio Lula da Silva").
- **Diferentes representações devido a caracteres especiais** – Violação de sintaxe, em virtude de diferentes formatos de representação dos dados não compostos, devido ao uso de caracteres especiais (e.g., uso de espaço, ausência de espaço, traço, parêntesis ou asterisco em campos de data, na dimensão $Tempo$).
- **Diferentes representações devido a ordenações diferentes** – Violação de sintaxe, em virtude de diferentes formatos de representação dos dados não compostos, devido ao uso de ordenações diferentes (e.g., uso de ordenações diferentes no campo $Nome$, como "Wesley Gongora" *vs* "Gongora, Wesley").
- **Diferentes representações devido a unidades de medida** – Violação de sintaxe, em virtude de diferentes formatos de representação dos dados não compostos, devido ao uso heterogeneidade nas unidades de medidas (e.g., campos com medidas diferentes para data, área, distância).
- **Diferentes representações devido a níveis de agregação** – Violação de sintaxe, em virtude de diferentes formatos de representação dos dados não compostos, devido ao uso de diferentes níveis de agregação (e.g., não observância do esquema de dados, ocasionando registros com diferentes agregações, como "ano-mês-dia" e "ano-trimeste-semana").
- **Valores *default* diferentes para dados ausentes** – O valor ausente é representado por diferentes valores *default* em uma tabela (e.g., a ausência de

registro no campo $Idade$ é preenchido com o valor 0 em alguns registros, e 1 em outros, em decorrência de problemas de integração dos dados, ocorridos na fase de ETL).

- **Violação de dependência funcional** – Uma dependência funcional envolvendo dois ou mais atributos no relacionamento dos dados é violada pela não obediência a cardinalidade dos dados (e.g., considerando a existência de uma dependência funcional entre os atributos CEP e $Bairro$, as tuplas $t1(72220060, 'Ceilândia')$ e $t2(72220060, 'Taguatinga')$ constituem uma violação de dependência funcional).
- **Violação de integridade referencial** – Existência dados órfãos ou pendurados (e.g., registros da Tabela Fato que apontam para outros registros que não existem em uma Dimensão).
- **Níveis de agregação hierarquizados incorretamente** – A modelagem da Dimensão apresenta níveis de agregação em hierarquia incorreta (e.g., Dimensão $Tempo$ apresenta os níveis em seqüência incoerente, como Ano, $Semana, Trimestre, Mês, Dia$).

4.2.6. Problemas na Dimensão da Acurácia

Nesta seção são apresentados os problemas de QD em DW que comprometem a acurácia dos dados em ambientes de DW. O significado de cada problema identificado é apresentado a seguir:

- **Presença de *outlier*** – Presença de valores atípicos, i.e., uma observação que apresenta um grande afastamento dos demais da série, ou que é inconsistente (e.g., considerando a existência de um campo *Idade*, a existência de um registro contendo valor igual a 140).
- **Entrada de dados errôneos** – Entrada de dados errôneos no DW, devido a inserção de valores incorretos na fonte de dados (e.g., campo $Idade$ contém atributo com o valor 25, quando na verdade deveria ser 28).
- **Dados com erros ortográficos** – Palavra escrita incorretamente devido a um erro ortográfico acidental (e.g., “Esteve Jób”, quando na verdade deveria ser “Steve Jobs”).

- **Entrada de dados estranhos** – Valor presente não corresponde ao tipo do registro (e.g., campo $Nome$ recebeu como entrada um valor do tipo numérico).
- **Violação do conjunto dados permitido** – Presença de valores não pertencentes ao conjunto de dados permitidos (e.g., campo $Cidade$ permite apenas o conjunto $\{Rio\ de\ Janeiro, Londres, Madrid\}$ e teve como entrada *"Brasília"*).
- **Referência incorreta** – Dimensão Fato aponta para o *ID* (ou identificador) errado, em relação à Dimensão (e.g., campo $ID\_localidade$ com valor 20, quando na verdade deveria ter recebido o valor 22 na tabela Fato).
- **Violação de restrição no relacionamento dos dados** – Conjunto de todos os problemas de violações no relacionamento dos dados, em função das regras de negócio (e.g., um empregado que tenha sido atribuído a um projeto, não lhe é permitido se inscrever em um programa de treinamento, i.e., supõe-se que os dados deste empregado não sejam encontrados na Dimensão $Treinamento$).

4.2.7. Resumo

O nível de granularidade no qual o problema ocorre é outra questão relevante na investigação dos problemas de QD desta proposta. Os modelos de dados podem ser classificados segundo a arquitetura que utilizam. Os dois principais são o modelo relacional, que surgiu para atender os sistemas transacionais, conhecidos como OLTP (do inglês *Online Transactional Processing*), e o modelo multidimensional, que surgiu com o propósito de atender os sistemas analíticos (OLAP). Uma das limitações das taxonomias de problemas de qualidade citadas na literatura reside em sua investigação centrada em modelos relacionais. A Figura 4.4 ilustra a estrutura de organização dos dados, segundo o modelo multidimensional.

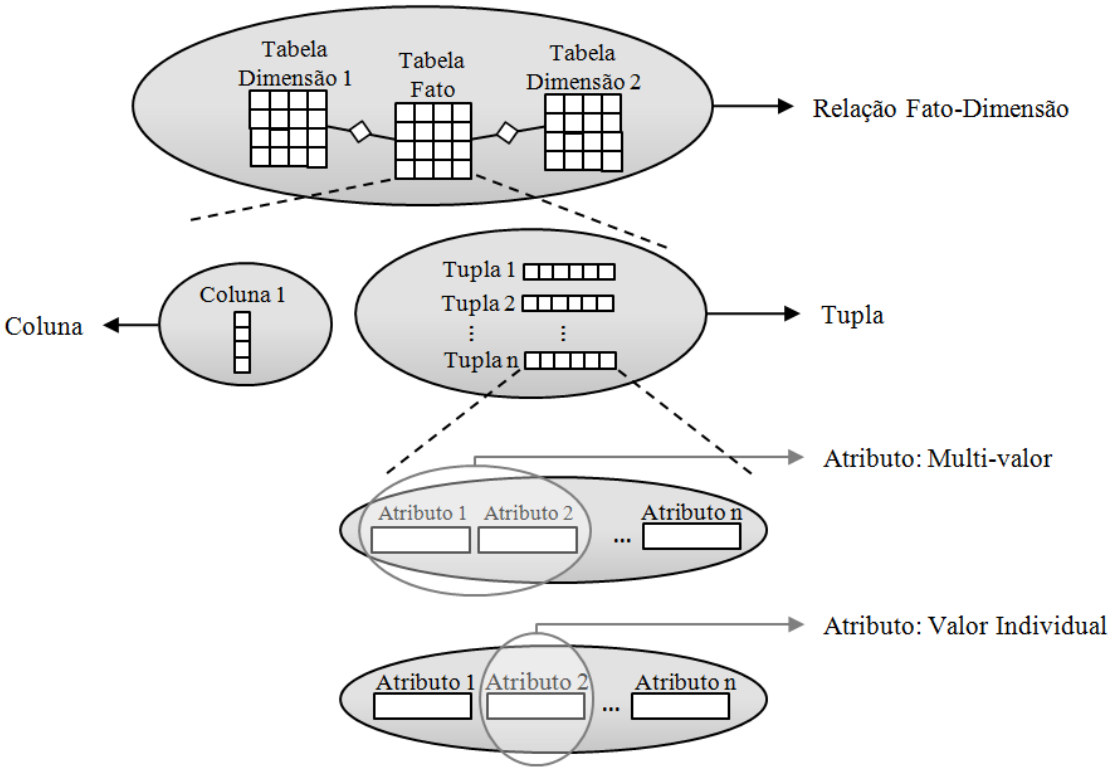


Figura 4.4: Estrutura de organização dos dados segundo o modelo multidimensional

O resumo da taxonomia dos problemas de QD em DW é apresentado na Tabela 4.1, contendo a dimensão, o problema de QD e o mapeamento em relação ao nível de granularidade no qual o problema se manifesta, i.e., valor individual, multi-valor, tupla, coluna ou relação Fato-Dimensão (ver Figura 4.4).

Tabela 4.1: Problemas de QD em DW

	No.	Problemas de DQ em DW	Atributo		Tupla	Coluna	Relação Fato-Dimensão
			Valor individual	Multi-valor			
COMPLETUDE	P.1	Valor ausente	X				
	P.2	Referência definida, mas não encontrada					X
	P.3	Purgar dados do DW	X	X	X	X	
	P.4	Ausência de lotes de dados			X		
ATUALIDADE	P.5	Ausência de atualizações periódicas			X	X	
	P.6	Intervalo de atualização incorreto			X	X	
	P.7	Dados desatualizados	X				

(continua)

	No.	Problemas de DQ em DW *(continuação)*	Atributo		Tupla	Coluna	Relação Fato-Dimensão
			Valor individual	Multi-valor			
UNICIDADE	P.8	Dados duplicados	X		X		
	P.9	Dados duplicados e contraditórios	X		X		
	P.10	Existência de sinônimos	X				
CONSISTÊNCIA	P.11	Modelagem do DW incompleta				X	
	P.12	Dimensão apontada por diferentes campos					X
	P.13	Diferentes tipos de dados para colunas semelhantes					X
	P.14	Diferentes representações devido a abreviações				X	X
	P.15	Diferentes representações devido a pseudômino/apelido				X	X
	P.16	Diferentes representações devido a caracteres especiais				X	X
	P.17	Diferentes representações devido a ordenações diferentes				X	X
	P.18	Diferentes representações devido a unidades de medida				X	X
	P.19	Diferentes representações devido a níveis de agregação				X	X
	P.20	Valores *default* diferentes para dados ausentes				X	X
	P.21	Violação de dependência funcional					X
	P.22	Violação de integridade referencial					X
	P.23	Níveis de agregação hierarquizados incorretamente				X	
ACURÁCIA	P.24	Presença de *outlier* (valores atípicos)	X				
	P.25	Entrada de dados errôneos	X				
	P.26	Dados com erros ortográficos	X				
	P.27	Entrada de dados estranhos *(Strings)*	X				
	P.28	Violação do conjunto de dados permitidos	X				
	P.29	Referência incorreta					X
	P.30	Violação de restrição no relacionamento dos dados					X

4.2.8. Comparação com trabalhos relacionados

A Tabela 4.2 apresenta uma relação de correspondência entre a taxonomia proposta nesta dissertação (i.e., três primeiras colunas) e as demais taxonomias para detecção dos problemas de qualidade de dados existentes na literatura [53, 54, 56, 67, 68] e apresentadas na seção 3.1.2, com as quais existe certa concordância na cobertura dos problemas de qualidade:

Tabela 4.2: Comparação com trabalhos relacionados

	No.	Problemas de DQ em DW	Rahm e Do, 2000	Müller e Freytag, 2003	Kim et al., 2003	Oliveira et al., 2005	Li et al., 2011
COMPLETUDE	P.1	Valor ausente	Valor ausente	Valor ausente	Valor ausente	Valor ausente	Valor ausente
	P.2	Referência definida, mas não encontrada	Violação integridade referencial	Violação integridade referencial	Violação restrição de integridade	Violação integridade referencial	Referência não encontrada
	P.3	Purgar dados do DW	–	–	–	–	–
	P.4	Ausência de lotes de dados	–	Tuplas ausentes	–	–	Tuplas ausentes
ATUALIDADE	P.5	Ausência de atualizações periódicas	–	–	–	–	–
	P.6	Intervalo de atualização incorreto	–	–	–	–	–
	P.7	Dados desatualizados	–	–	Dados temporais desatualizados	–	Valores desatualizados
UNICIDADE	P.8	Dados duplicados	Violação de unicidade	Violação restrição de integridade	Dados duplicados	Violação de unicidade	Registro duplicado
	P.9	Dados duplicados e contraditórios	–	–	–	Duplicados inconsistentes	–
	P.10	Existência de sinônimos	Diferentes representações dos valores	Irregularidade	Valores ambíguos	Existência de sinônimos	Valores ambíguos / abreviaturas
CONSISTÊNCIA	P.11	Modelagem do DW incompleta	Desenho do esquema pobre	–	–	–	–
	P.12	Dimensão apontada por diferentes campos	–	–	–	–	–
	P.13	Diferentes tipos de dados para colunas semelhantes	–	–	–	–	–
	P.14	Diferentes representações devido a abreviações	Transposição de valores	Erro sintático	Dif. repre-sentações, devido a abreviações	Violação de sintaxe	Dif. repre-sentações, devido a abreviações
	P.15	Diferentes representações devido à pseudômino/apelido	-	Erro sintático	Dif. repre-sentações, devido a pseudônimo / apelido	Violação de sintaxe	Violação de sintaxe
	P.16	Diferentes representações devido a caracteres especiais	-	Erro sintático	Dif. repre-sentações, devido a caracteres especiais	Violação de sintaxe	Dif. repre-sentações, devido a caracteres especiais

(continua)

	No.	Problemas de DQ em DW *(continuação)*	Rahm e Do, 2000	Müller e Freytag, 2003	Kim et al., 2003	Oliveira et al., 2005	Li et al., 2011
CONSISTÊNCIA	P.17	Diferentes representações devido a ordenações diferentes	Transposição de palavras	Erro sintático	Diferentes representações, devido a ordenações diferentes	Violação de sintaxe	Dif. representações, devido a ordenações diferentes
	P.18	Diferentes representações devido a unidades de medida	Dados inconsistentes	Erro sintático	Diferentes representações, devido a unidades medida	Violação de sintaxe	Dif. representações, devido a unidades medida
	P.19	Diferentes representações devido a níveis de agregação	Dados inconsistentes	Erro sintático	Diferentes representações, devido a níveis agregação	Violação de sintaxe	Dif. representações, devido a níveis agregação
	P.20	Valores *default* diferentes para dados ausentes	–	–	–	–	–
	P.21	Violação de dependência funcional	Violação de depend. entre atributos	Violação de restrição de integridade	Valores mutuamente inconsistentes	Violação de dependência funcional	Relacionamento de cardinalidade
	P.22	Violação de integridade referencial	Violação integridade referencial	Violação integridade referencial	Violação restrição de integridade	Violação integridade referencial	Referência não encontrada
	P.23	Níveis de agregação hierarquizados incorretamente	–	–	–	–	–
ACURÁCIA	P.24	Presença de *outlier* (i.e, valores numéricos atípicos)	Genericamente, valor ilegal	–	Violação do intervalo de valores	Genericamente, violação domínio	Valores fora faixa
	P.25	Entrada de dados errôneos	–	–	Entrada de dados errôneos	–	Entradas errôneas
	P.26	Dados com erros ortográficos	Erro ortográfico	–	Erro ortográfico	Erro ortográfico	Erro ortográfico
	P.27	Entrada de dados estranhos *(Strings)*	Genericamente, valor ilegal	–	Dados estranhos	Genericamente, violação domínio	Dados estranhos
	P.28	Violação do conjunto de dados permitidos	Genericamente, valor ilegal	Genericamente, violação restrição de integridade	Violação do intervalo de valores	Genericamente, violação domínio	Violação do conjunto de dados permitidos
	P.29	Referência incorreta	–	–	–	–	–
	P.30	Violação de restrição no relacionamento dos dados	–	–	–	–	Violação de restrição no relacionamento dos dados

Ao observar as abordagens adotadas em outras taxonomias, a proposta de detecção *bottom-up* a partir dos níveis de granularidade, adotada por Oliveira et al. [54] mostrou-se mais adequada para o mapeamento dos problemas de qualidade. Entretanto, adotá-la como critério de classificação não foi possível, já que muitos problemas manifestam-se em mais de um nível. Para este propósito, a organização dos itens em dimensões, como em Li et al. [56], mostrou-se mais satisfatória. Através da comparação entre os problemas de QD das taxonomias relacionadas é possível constatar o seguinte:

- Os problemas de QD manifestos em bancos de dados transacionais divergem dos modelos multidimensionais, característicos de DW, mesmo existindo certa relação em alguns itens. Por exemplo, nos problemas de QD citados por Oliveira et al. [54], a subcategoria *Problemas ao nível de múltiplas fontes de dados* foi desconsiderado em nossa compilação. Entretanto, caso o objeto de estudo fosse a integração de dados, seu conjunto de problemas seria de grande valor;

- Embora o número de problemas de qualidade contidos em nossa taxonomia seja inferior quantitativamente em relação aos trabalhos de outros autores (i.e., como a Taxonomia de Dados Sujos [53] e a Taxonomia de Dados Sujos Baseados em Regras [56]), a investigação apresentada neste trabalho é mais abrangente e oferece uma cobertura mais adequada para os problemas de QD manifestas em ambientes de DW/BI;

- Todos os diferentes problemas identificados em trabalhos anteriores e que podem ser aplicados em DW também têm suporte na nossa taxonomia, embora os nomes possam ser diferentes. Na taxonomia proposta, procurou-se que as designações estivessem de acordo com a terminologia comum a ambientes de *Data Warehouse* e de *Business Intelligence* (e.g., o problema *P.3 Purgar dados do DW* corresponde a uma situação exclusiva do domínio investigado neste trabalho).

- Muitos problemas mapeados nesta nova proposta de taxonomia já haviam sido identificados em trabalhos anteriores. Entretanto, seu nível de cobertura era parcial para o domínio de aplicações de DW/BI. Os problemas classificados nesta condição encontram-se identificados através de células hachuradas, nas cinco últimas colunas da Tabela 4.2.

- Através do quadro comparativo, é possível observar que nossa taxonomia é mais apropriada para o nosso propósito inicial. A taxonomia proposta apresenta oito novos problemas de QD em DW que não constam em nenhuma das demais taxonomias relacionadas, especificamente: *P3. Purgar dados do DW*, *P5. Ausência de atualizações periódicas*, *P6. Intervalo de atualização incorreto*, *P12. Dimensão apontada por diferentes campos*, *P13. Diferentes tipos de dados para colunas semelhantes*, *P20. Valores default diferentes para dados*

ausentes, *P23. Níveis de agregação hierarquizados incorretamente* e *P29. Referência incorreta*.

O processo de concepção da taxonomia partiu de uma revisão exaustiva e sistemática dos problemas de qualidade. Ainda assim, não é possível provar formalmente que ela completa, i.e., que cobre garantidamente todos os problemas que afetam os dados em ambientes de DW/BI. No entanto, a comparação com os trabalhos relacionados evidencia que os problemas identificados em trabalhos anteriores possuem correspondente na taxonomia proposta. Por outro lado, a taxonomia proposta apresenta novos problemas de QD que não possuem correspondência em nenhuma dessas taxonomias. Além disso, cada uma das 117 causas de problemas de QD em *Data Warehouses* apresentados por Singh et al. (2011) [55] e as 13 categorias de processos que causam problemas de qualidade identificadas por Maydanchik (2007) [71] foram cuidadosamente analisadas, e transformadas em itens que afetam a qualidade em neste trabalho. Estas constatações contribuem para a convicção de que se trata de uma taxonomia mais completa e adequada para tratar os problemas de qualidade que afetam modelos multidimensionais em ambientes de DW/BI, quando comparado com as propostas já existentes.

4.3. MÉTRICAS OBJETIVAS PARA A QUALIDADE DOS DADOS

Nesta seção, são apresentadas as métricas que foram propostas para a qualidade/confiabilidade dos dados em ambientes de DW/BI. Conforme discutido no Capítulo 3, a confiabilidade tem sido freqüentemente citada na literatura. Entretanto, poucos trabalhos tem se dedicado a analisar mais profundamente a questão. Como medidas de QD são freqüentemente desenvolvidas numa base *ad hoc* para resolver problemas específicos e práticos [40], elas geralmente contêm um alto grau de subjetividade [31]. Essa lacuna é suprida nesta investigação através da elaboração de métricas prioritariamente objetivas através dos problemas descobertos na Seção 4.2.

4.3.1. Requisitos das métricas de QD

Para garantir uma base científica e permitir o desenvolvimento de métricas aplicáveis e bem-fundamentadas, os Requisitos para Métricas de DQ propostos por Heinrich et al., (2007; 2011) [57, 58] foram utilizados como referência. Elas foram

derivadas de seis requisitos para métricas de qualidade da literatura [40, 59, 60, 61] e podem ser resumidas como segue:

R1. [*Normalização*] Uma normalização adequada é necessária para permitir a comparação dos valores das métricas (e.g., para comparar os diferentes níveis de QD ao longo do tempo). Neste contexto, as métricas de QD são tipicamente proporções com um valor entre 0 (fraco) e 1 (perfeito).

R2. [*Escala intervalar*] Para suportar tanto o monitoramento das mudanças no nível de QD ao longo do tempo e a avaliação econômica das medidas, é necessário que as métricas estejam em uma escala intervalar.

R3. [*Interpretabilidade*] A quantificação precisa ser de "fácil interpretação pelos usuários de negócio". Por essa razão, as métricas de QD têm de ser compreensíveis.

R4. [*Agregação*] A QD tem de ser quantificada em níveis diferentes. Para um modelo de dados relacional, isso implica a capacidade de quantificar QD sobre o nível de valores de atributo, tuplas e relações, bem como sobre o banco de dados inteiro para que os valores tenham interpretação semântica consistente em cada nível. Além disso, as métricas devem permitir a agregação dos resultados quantificados para o nível mais elevado subseqüente.

R5. [*Adaptatividade*] Para quantificar a QD de uma forma orientada a objetivos (do inglês g*oal-oriented*), as métricas devem ser adaptáveis ao contexto de uma aplicação particular.

R6. [*Viabilidade*] Para assegurar a aplicabilidade, as métricas precisam ser baseadas em parâmetros de entrada mensuráveis. Ao definir as métricas, métodos de medição devem ser definidos e, onde a medição exata não é possível ou de custo muito elevado, métodos alternativos (e.g., estatística) têm de ser propostos.

Para suportar os Requisitos, um conjunto de formas funcionais propostas na literatura [40, 79] para o cálculo dos valores das métricas foi revisado e selecionado. Sua adoção encontra-se em conformidade com o objetivo de definir métricas aplicáveis e bem-fundamentadas, conforme requisitos de Heinrich et al., (2007; 2011) [57, 58]:

- **Razão simples (*Simple ratio*)** – Razão entre o número de resultados necessários para o número total de resultados. Sua forma preferida de utilização dá-se pelo número de resultados indesejáveis dividido pelo total de resultados subtraídos de 1.

- **Operação min ou max (*Min or max operation*)** – Para lidar com dimensões que requerem a agregação de vários indicadores de QD (variáveis), a operação mínima (min) e máxima (max) pode ser aplicada entre os valores normalizados. O operador min é conservador na medida em que atribui à métrica o valor mais pessimista dentre todos os valores agregados. A operação max é usada se uma interpretação liberal é garantida. As variáveis individuais podem ser medidas utilizando a razão simples.

- **Média ponderada (*Weighted average*)** – Média ponderada de métricas de valores para diferentes atributos ou objetos de dados. Se a importância de cada métrica seja bem compreendida para avaliação de uma dimensão, a utilização da média ponderada será apropriada.

- **Booleano (*Boolean*)** – É representado por um valor Booleano. Valor igual a 1 se a condição é verdadeira $(true)$, e caso contrário falso $(false)$, se valor igual a 0.

- **Grau (*degree*)** – Grau de probabilidade com que determinado evento seja verdadeiro. Tal grau é comumente representado pelo intervalo [0-1].

A partir da Tabela 4.1, cada problema de QD foi atentamente analisado e transformado em uma métrica objetiva. Esse processo de construção da solução *bottom-up* permite o cumprimento dos requisitos para Métricas de QD. Por exemplo, para o Requisito Agregação, uma métrica de nível igual a $n + 1$ (e.g., completude sobre o nível de tuplas) baseia-se na métrica correspondente ao nível n (e.g., completude sobre o nível de valores de atributo).

Um Glossário para as Notações Utilizadas para as métricas de QD propostas é apresentado na Tabela 4.3.

Tabela 4.3: Glossário das Notações Utilizadas

Símbolo	Definição
N	O número total de campos da tupla
C_i	Campo do banco de dados
$v(C_i)$	Representa o valor v de um campo C_i. Onde $v(C_i)$ é 1 se o campo C_i tem um valor não-nulo, 0 caso contrário
N_i	O número total de tuplas
$v_e(i)$	Valores errôneos
$v_a(i)$	Valores atípicos (ou *outliers*)

Como as métricas encontram-se classificadas em suas respectivas dimensões, elas serão apresentadas separadamente. Durante seu processo de elaboração, preocupou-se em estabelecer métricas quantitativas que permitam avaliação objetiva e automática, levando a custos mais baixos de medição, especialmente no caso de conjuntos enormes de dados [57], i.e., ambientes de BI. Formalmente, cada problema de QD pode ser avaliado por – pelo menos – uma métrica correspondente, conforme apresentado nas seções seguintes.

4.3.2. Métricas Objetivas para a Completude

Completude é o grau no qual os elementos de um esquema estão presentes nas instâncias. Dado o conjunto dos problemas de completude dos dados $P1, P2, P3, P4$, conforme Seção 4.2.2, quatro métricas correspondentes são propostas para aferir a completude. Na Tabela 4.4 as métricas são apresentadas.

Tabela 4.4: Métricas para a Completude dos Dados

Problema de QD	Métrica	Forma Funcional	Cálculo da Métrica
P1	% dos registros onde o campo está preenchido ou está identificado como não aplicável	Razão Simples	$Completude_{m1=} \frac{\sum_{i=1}^{N} v(C_i)}{N}$
P2	Número de dados órfãos ou pendurados	Booleano	$VERDADEIRO$, se a referência definida existe, caso contrário, $FALSO$
P3	% de registros que foram purgados em etapas anteriores a carga no DW	Razão Simples	$Completude_{m3=} \frac{\sum_{i=1}^{N} C_i}{N}$
P4	% de tuplas de um lote de carga que estão ausentes no DW	Razão Simples	$Completude_{m4=} \frac{\sum_{i=1}^{N} T_i}{N_t}$

4.3.3. Métricas Objetivas para a Atualidade

A partir dos três problemas de atualidade dos dados, conforme investigação da Seção 4.2.3, apresentamos na Tabela 4.5 as métricas correspondentes aos problemas $P5, P6$ e $P7$ de QD:

Tabela 4.5: Métricas para a Atualidade dos Dados

Problema de QD	Métrica	Forma Funcional	Cálculo da Métrica
P5	Ocorrência de atualizações periódicas	Booleano	$VERDADEIRO$, se a atualização está ocorrendo periodicamente, caso contrário, $FALSO$
P6	Intervalo de atualizações do DW	Booleano	$VERDADEIRO$, se a atualização está occorendo no intervalo esperado, caso contrário, $FALSO$
P7	% de registros desatualizados (não foram atualizados desde a época de extração) i.e., *delay* entre a sua alteração na fonte e a replicação no DW	Max	$max\left(0; 1 - \frac{Valor\ Corrente}{Volatilidade}\right)$

4.3.4. Métricas Objetivas para a Unicidade

Nesta seção são apresentadas as métricas para a dimensão da unicidade dos dados. Conforme investigação dos problemas de QD da Seção 4.2.4, apresentamos na Tabela 4.6 as métricas correspondentes aos problemas de QD $P8, P9$ e $P10$:

Tabela 4.6: Métricas para a Unicidade dos Dados

Problema de QD	Métrica	Forma Funcional	Cálculo da Métrica
P8	Número de dados duplicados	Booleano	$VERDADEIRO$, se a tupla apresenta alta probabilidade de duplicidade, caso contrário, $FALSO$
P9	Número de dados duplicados e contraditórios	Booleano	$VERDADEIRO$, se a tupla apresenta alta probabilidade de duplicidade contraditória, caso contrário, $FALSO$
P10	Número de valores de dados de significado idêntico ou muito semelhante ao de outros	Booleano	$VERDADEIRO$, se existe sinônimo, caso contrário, $FALSO$

4.3.5. Métricas Objetivas para a Consistência

Nesta seção são apresentadas as métricas para a dimensão da consistência dos dados. Conforme investigação dos problemas de QD da seção 4.2.5, apresentamos na Tabela 4.7 as métricas correspondentes aos problemas de QD $P11$ ao $P23$:

Tabela 4.7: Métricas para a Consistência dos Dados

Problema de QD	Métrica	Forma Funcional	Cálculo da Métrica
P.11	Número de colunas ausentes na modelagem do DW	Max	$max\left(0; \sum colunas\ ausentes\right)$
P.12	Número de *Primary Keys (PK)* de uma tabela	Max	$max\left(1; \sum PK\ de\ uma\ dimensão\right)$
P.13	Ocorrência de erros na definição de tipos	Booleano	$VERDADEIRO$, se a coluna é definida e referenciada por diferentes tipos na modelagem, caso contrário, $FALSO$
P.14	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a abreviações	Max	$max\left(1; \sum rep_{abrev}\right)$
P.15	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a pseudônimo/apelido	Max	$max\left(1; \sum rep_{apelido}\right)$
P.16	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a caracteres especiais	Max	$max\left(1; \sum rep_{caract.}\right)$
P.17	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a ordenações diferentes	Max	$max\left(1; \sum rep_{orden}\right)$
P.18	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a unidades de medida	Max	$max\left(1; \sum rep_{und}\right)$
P.19	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a níveis de agregação diferentes	Max	$max\left(1; \sum rep_{agreg}\right)$
P.20	Número de valores *default* para dados ausentes na tabela (Dimensão ou Fato)	Max	$max\left(1; \sum valores\ \ default\right)$
P.21	Ocorrência de registros com violação de dependência funcional	Max	$VERDADEIRO$, se ocorre violação da dependência funcional, caso contrário, $FALSO$
P.22	Ocorrência de violações de integridade referencial	Booleano	$VERDADEIRO$, se ocorre violação da integridade referencial, caso contrário, $FALSO$
P.23	Ocorrência de ordenação hierárquica incorreta	Booleano	$VERDADEIRO$, ordenação hierárquica incorreta, caso contrário, $FALSO$

4.3.6. Métricas Objetivas para a Acurácia

Nesta seção são apresentadas as métricas para a dimensão da acurácia dos dados. Conforme investigação dos problemas de QD da seção 4.2.6, apresentamos na Tabela 4.8 as métricas correspondentes aos problemas de QD $P24$ ao $P30$:

Tabela 4.8: Métricas para a Acurácia dos dados

Problema de QD	Métrica	Forma Funcional	Cálculo da Métrica
P.24	% de valores atípicos (ou *outliers)*	Razão Simples	$Acurácia_{m24=} \frac{\sum_{i=1}^{N} v_a(C_i)}{N}$
P.25	% de valores errôneos	Razão Simples e grau	$Acurácia_{m25=} \frac{\sum_{i=1}^{N} v_e(i)}{N}$
P.26	Número de registros com erro ortográfico	Max	$max\left(0; \sum erro\ ortográfico\right)$
P.27	Número de registros estranhos	Max	$max\left(0; \sum dados\ estranhos\right)$
P.28	Número de registros que violam o conjunto de dados permitidos	Max	$max\left(0; \sum viola\ dados\ perm.\right)$
P.29	Número de registros referência incorreta	Max	$max\left(0; \sum ref.incorreta\right)$
P.30	Número de registros que violam a restrição no relacionamento dos dados	Booleano	$VERDADEIRO$, se registro viola a restrição no relacionamento dos dados, caso contrário, $FALSO$

4.3.7. Dificuldade de Medição

As dimensões de QD possuem diferentes níveis de dificuldade de medição e de impacto no negócio. Graças à lista de métricas, esse fenômeno pode ser mais bem compreendido (ver Figura 4.5).

O impacto no negócio está relacionado com a dificuldade de medição em uma relação de proporcionalidade direta, ou seja, à medida que a sua complexidade de medição aumenta, seu impacto no negócio cresce na mesma proporcionalidade. Este fato deve-se a dificuldade e, em alguns casos, a impossibilidade de transformar problemas em métricas ou em métodos práticos de aferição objetiva. Esse problema em especial, refere-se à

dimensão da acurácia que tem sido, desde o início das pesquisas de QD até a presente data, a dimensão mais citada na literatura [27, 86].

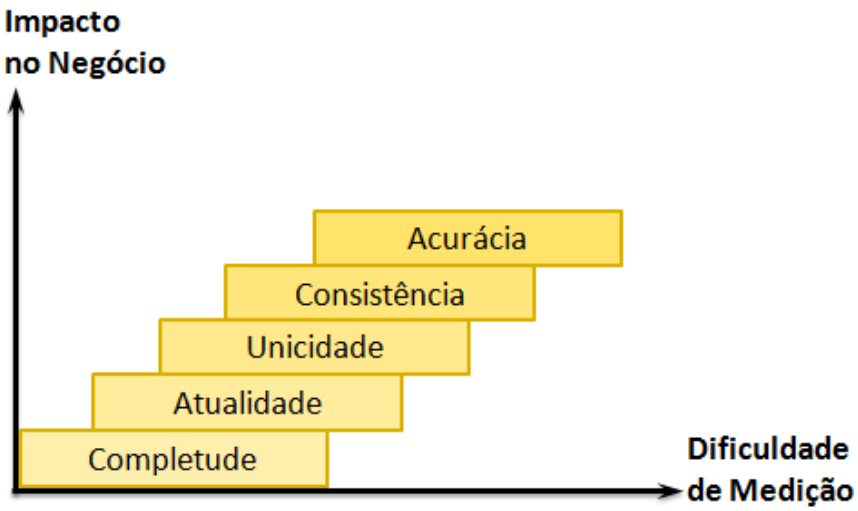


Figura 4.5: Dificuldade de Medição das dimensões de QD

Obter metadados de qualidade que reflitam o real estado dos dados, de forma consistente e precisa, não é tarefa fácil. Entretanto, este trabalho baseia-se no princípio que, somente a partir de métricas alinhadas aos problemas de QD e/ou dados sujos, avaliações confiáveis possam ser mais bem asseguradas.

4.3.8. Resumo das Métricas

O resumo das métricas para avaliação da confiabilidade/qualidade dos dados em DW/BI é apresentado na Tabela 4.9. O nível de subjetividade é identificado e está diretamente ligada a Figura 4.5, que ilustra a dificuldade no processo de medição. É apresentado ainda, na última coluna, um comparativo com trabalhos existentes na literatura sobre métricas de QD.

O resultado desta análise mostra claramente, não só a impossibilidade de manter ambientes de DW/BI livres de erros, mas a natureza incerta quando a questão se refere a dados 100% confiáveis:

Tabela 4.9: Métricas de Confiabilidade dos Dados

Problema de QD		Métrica	Forma Funcional	Cálculo da Métrica	Existência de Subjetividade *(complexidade)*	Citada por
COMPLETUDE	P1	% dos registros onde o campo está preenchido ou está identificado como não aplicável	Razão Simples	$C_{m1=} \frac{\sum_{i=1}^{N} v(C_i)}{N}$		[28, 80]
	P2	Número de dados órfãos ou pendurados	Booleano	$VERDADEIRO$, se a referência definida existe, caso contrário, $FALSO$		
	P3	% de registros que foram purgados em etapas anteriores a carga no DW	Razão Simples	$C_{m3=} \frac{\sum_{i=1}^{N} C_i}{N}$		
	P4	% de tuplas de um lote de carga que estão ausentes no DW	Razão Simples	$C_{m4=} \frac{\sum_{i=1}^{N} T_i}{N_t}$		
ATUALIDADE	P5	Ocorrência de atualizações periódicas	Booleano	$VERDADEIRO$, se a atualização está ocorrendo periodicamente, caso contrário, $FALSO$		
	P6	Intervalo de atualizações do DW	Booleano	$VERDADEIRO$, se a atualização está occorendo no intervalo esperado, caso contrário, $FALSO$		
	P7	% de registros desatualizados (não foram atualizados desde a época de extração) i.e., *delay* entre a alteração na fonte e sua a replicação no DW	Max	$max\left(0; 1 - \frac{Valor\ corrente}{Volatilidade}\right)$		[28, 58, 79, 81]
UNICIDADE	P8	Número de dados duplicados	Booleano	$VERDADEIRO$, se a tupla apresenta alta probabilidade de duplicidade, caso contrário, $FALSO$		[28, 80]
	P9	Número de dados duplicados e contraditórios	Booleano	$VERDADEIRO$, se a tupla apresenta alta probabilidade de duplicidade contraditória, caso contrário, $FALSO$		
	P10	Número de valores de dados de significado idêntico ou muito semelhante ao de outros	Booleano	$VERDADEIRO$, se existe sinônimo, caso contrário, $FALSO$	Alta	
CONSISTÊNCIA	P.11	Número de colunas ausentes na modelagem do DW	Max	$max\left(0; \sum colunas\ ausentes\right)$	Alta	[80]
	P.12	Número de *Primary Keys (PK)* de uma tabela	Max	$max\left(1; \sum PK\ da\ dimensão\right)$		
	P.13	Ocorrência de erros na definição de tipos	Booleano	$VERDADEIRO$, se a coluna é definida e referenciada por diferentes tipos na modelagem, caso contrário, $FALSO$		

(continua)

Problema de QD		Métrica *(continuação)*	Forma Funcional	Cálculo da Métrica	Existência de Subjetividade *(complexidade)*	Citada por
CONSISTÊNCIA	P.14	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a abreviações	Max	$max\left(1; \sum rep_{abrev}\right)$	Média	[28, 79]*
	P.15	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a pseudônimo/apelido	Max	$max\left(1; \sum rep_{apelido}\right)$	Média	[28, 79]*
	P.16	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a caracteres especiais	Max	$max\left(1; \sum rep_{caract.}\right)$	Média	[28, 79]*
	P.17	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a ordenações diferentes	Max	$max\left(1; \sum rep_{orden}\right)$	Média	[28, 79]*
	P.18	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a unidades de medida	Max	$max\left(1; \sum rep_{und}\right)$	Média	[28, 79]*
	P.19	Número de registros com formato de representação inconsistente na tabela (Dimensão ou Fato) devido a níveis de agregação diferentes	Max	$max\left(1; \sum rep_{agreg}\right)$	Média	[28, 79]*
	P.20	Número de valores *default* para dados ausentes na tabela (Dimensão ou Fato)	Max	$max\left(1; \sum \begin{matrix} valores \\ default \end{matrix}\right)$		
	P.21	Ocorrência de registros com violação de dependência funcional	Max	$VERDADEIRO$, se ocorre violação da dependência funcional, caso contrário, $FALSO$		
	P.22	Ocorrência de violações de integridade referencial	Booleano	$VERDADEIRO$, se ocorre violação da integridade referencial, caso contrário, $FALSO$		
	P.23	Ocorrência de ordenação hierárquica incorreta	Booleano	$VERDADEIRO$, ordenação hierárquica incorreta, caso contrário, $FALSO$	Alta	

(continua)

Problema de QD		Métrica *(continuação)*	Forma Funcional	Cálculo da Métrica	Existência de Subjetividade *(complexidade)*	Citada por
ACURÁCIA	P.24	% de valores atípicos (ou *outliers)*	Razão Simples	$A_{m19=} \frac{\sum_{i=1}^{N} outlier}{N}$		
	P.25	% de valores errôneos	Razão Simples e grau	$A_{m20=} \frac{\sum_{i=1}^{N} v_e(i)}{N}$		
	P.26	Número de registros com erro ortográfico	Max	$max\left(0; \sum erro\ ortográfico\right)$	Alta e probabilística	
	P.27	Número de registros estranhos	Max	$max\left(0; \sum dados\ estranhos\right)$	Alta e probabilística	[79]
	P.28	Número de registros que violam o conjunto de dados permitidos	Max	$max\left(0; \sum viola\ dados\ perm.\right)$		
	P.29	Número de registros referência incorreta	Max	$max\left(0; \sum ref.incorreta\right)$	Alta e probabilística	
	P.30	Número de registros que violam a restrição no relacionamento dos dados	Booleano	$VERDADEIRO$, se registro viola a restrição no relacionamento dos dados, caso contrário, $FALSO$	Alta	

** Trabalhos no qual a métrica para a consistência é citada de forma genérica.*

Além da impossibilidade de manter ambientes de DW/BI livres de erros e a natureza incerta quando a questão se refere a dados 100% confiáveis, é evidente também a incapacidade de avaliação de todas as características (ou problemas) que interferem na QD. Isso ocorre em conseqüência de alguns tipos de problemas/métricas dificilmente serem convertidas em métodos eficazes e objetivos (i.e, ausência de subjetividade) de aferição, salvo pelo uso de metadados, soluções heurísticas e/ou tratamento manual. Outros problemas residem em sua natureza altamente contextual (e.g. dependentes de contexto de uso).

Embora todos os problemas de QD tenham sido convertidos em métricas, o mapeamento destas métricas em métodos eficazes me medição nem sempre será possível, pelo menos através de métodos convencionais. Um exemplo de problema que apresenta esta dificuldade é o *P.25 - Entradas errôneas*. Ao observar registros com esse problema, é possível constatar que eles pertencem ao domínio de valores válidos e não apresentam discrepâncias capazes de detecção via técnicas de *outliers* (ou valores atípicos). Essa ocorrência pode ser exemplificada: em um campo tipo numérico que armazene a $Idade$, o

valor do registro inserido foi "27", ao invés de "25". Esse simples exemplo demonstra que é impraticável assumir a existência de dados totalmente confiáveis. Essa observação confirma a questão da incerteza inerente aos dados.

Por fim, é possível concluir que a proposta apresentada neste trabalho é mais abrangente e oferece uma cobertura mais adequada para as métricas de QD obtidas a partir do conjunto de problemas de QD em ambientes de *Data Warehouse* e *Business Intelligence.* Devido ao grande número de métricas citados na literatura, apenas trabalhos voltados a consolidação de métricas de QD foram considerados no comparativo com trabalhos relacionados (referenciados na última coluna da Tabela 4.9).

4.4. AGREGAÇÃO DAS MÉTRICAS

A agregação é um processo utilizado em muitas tecnologias, especialmente na tomada de decisão multicriterial.

Formalmente, assuma D_i sendo uma dimensão com as métricas $D_i.m_1$, $D_i.m_2$, $\dots, D_i.m_n$ para os valores das avaliações das métricas m_1, m_2, $\dots$, m_n. A importância relativa de uma métrica $m_{i,}$ no que diz respeito à dimensão D_i, pode ser ponderada com um peso não-negativo $\alpha_i \in [0; 1]$. Conseqüentemente, a pontuação geral sobre a avaliação de uma dimensão pode ser definida pela Equação 4.1:

$$D_i = \sum_{i=1}^{n} \frac{m_i \,.\, \alpha_i}{n} \qquad \text{(Eq. 4.1)}$$

A agregação deve resultar em uma pontuação entre 0 e 1 que não pode ser superior ao nível da mais alta qualidade, ou menor do que a mais baixa, entre os itens granular. Esse requisito é fundamental para a consistência de agregação [57].

Portanto, quando todos os itens granulares são de qualidade idêntica, a agregação deve resultar na mesma pontuação. O operador de agregação aqui utilizado é a média ponderada [40], onde os pesos são as atribuições de valor embutido.

4.5. MODELO FUZZY DE CONFIABILIDADE DE DADOS

Sistemas baseados em lógica difusa (ou *fuzzy*) têm como objetivo alcançar robustez, tratabilidade e baixo custo sem uma modelagem matemática. Ao invés de lidar com uma formulação matemática do processo de agregação das métricas, imita-se o especialista.

Além do mais, empregou-se a lógica *fuzzy*, porque ela tem a capacidade de simular o raciocínio humano e tem uma importante vantagem sobre *hard-computing* quando se lida com comandos não exatos. No contexto tratado, precisou-se da lógica *fuzzy* porque esta pode lidar com a imprecisão que está envolvida na avaliação da confiabilidade dos dados. Na Figura 4.6 ilustra-se o processo *fuzzy* de avaliação da confiabilidade dos dados em ambientes de BI.

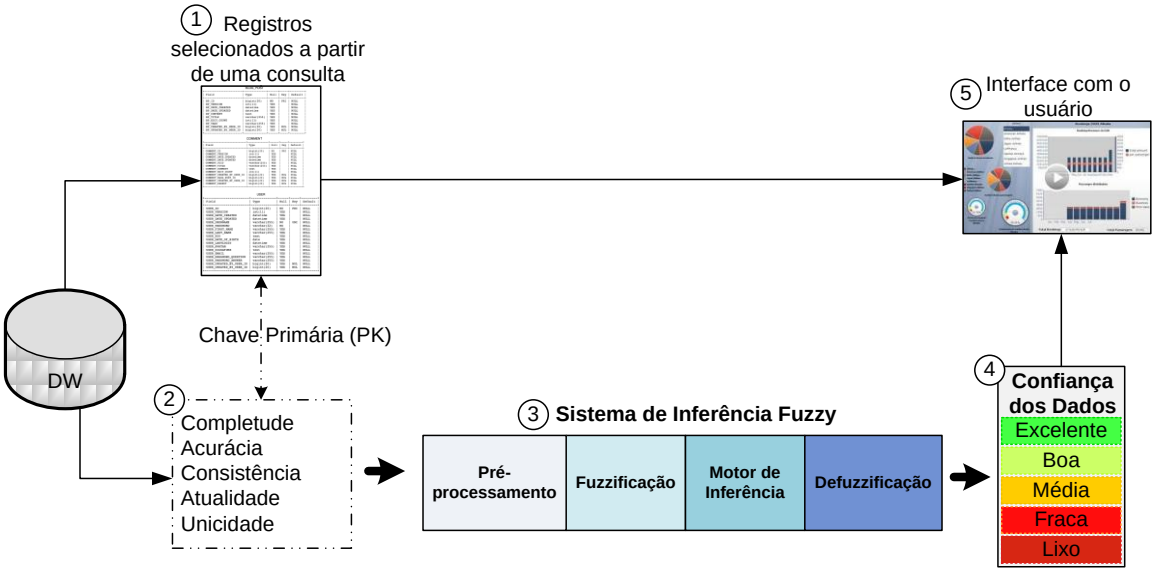


Figura 4.6: Processo fuzzy de avaliação da confiabilidade dos dados

O sistema de inferência *fuzzy* é composto de cinco etapas que são inter-relacionadas:

1. **Registros Selecionados:** O processo de inferência inicia-se com uma consulta simples e regular ao repositório de dados (DW). Os registros solicitados através da interface do BI são disponibilizados ao cliente sem nenhuma alteração. Entretanto, essa ação desencadeia um processo paralelo não perceptível ao usuário;
2. **Metadados de Confiabilidade dos Dados:** É responsável por manter os metadados materializados em um modelo multidimensional. Sua operação

consiste em receber o valor da Chave Primária (PK) corresponde ao registro(s) selecionado(s). Os metadados referentes à Completude, Acurácia, Consistência, Atualidade e Unicidade são então selecionados e introduzidos no motor de inferência *fuzzy*. A manutenção do repositório depende de regras de negócio corporativas e, de acordo com a freqüência escolhida (diária, semanal, mensal, etc), o sistema de BI realiza atualizações incrementais em sua base.

3. **Sistema de Inferência *Fuzzy*:** É o componente central do processo. É representado por uma máquina de inferência que faz uso de um conjunto de regras *fuzzy* criadas previamente. À medida que os dados são solicitados no sistema de *BI*, a máquina recebe as suas pontuações relativas à confiabilidade.
4. **Confiabilidade dos Dados (*defuzzification*):** Disponibilização da pontuação agregada das dimensões de confiabilidade dos dados. As seguintes variáveis lingüísticas foram adotadas $\{Lixo, Fraca, Média, Boa, Excelente\}$.
5. **Interface com o Usuário:** À medida que os dados vão sendo manipulados e exibidos no sistema de BI para o usuário, a solução proposta permite que o nível de confiabilidade dos dados seja exibido conjuntamente.

Através do valor de qualidade é possível verificar o quão confiáveis são os dados, utilizando-se funções trapezoidais para simular o comportamento do especialista em qualidade dos dados.

O modelo Mamdani [75] foi adotado para o sistema de inferência *fuzzy* e apresenta as seguintes características:

- Método de cálculo de *AND*: mínimo;
- Método de cálculo de *OR*: máximo;
- Método de cálculo de implicação: mínimo;
- Método de cálculo de agregação: máximo;
- Método de *defuzzificação*: centróide (ou Centro de Gravidade);

Na escolha de um *defuzzificador*, os critérios de plausividade (o valor de saída é intuitivo), simplicidade computacional, e de continuidade foram considerados.

4.5.1. Função de fuzzificação para a variável de entrada Completude

Os conjuntos *fuzzy* para a dimensão da completude apresentam três funções de pertinência trapezoidais e têm como limite inferior -0.1 e limite superior 1.2. A Figura 4.7 ilustra o conjunto *fuzzy* gerado para a variável. Definiu-se que os valores lingüísticos possíveis são $\{fraca, boa$ e $excelente\}$.

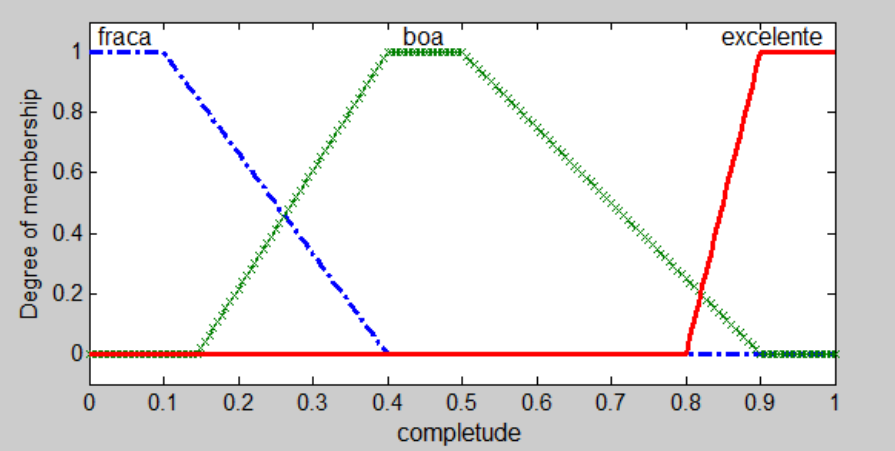


Figura 4.7: Funções de Pertinência da variável de entrada Completude

4.5.2. Função de fuzzificação para a variável de entrada Acurácia

O mesmo se aplica para a variável de entrada acurácia, que também é composta por três funções de pertinência (duas trapezoidais e uma triangular) e seus conjuntos fuzzy apresentam a seguinte característica: limite inferior -0.1 e limite superior 1.2. A Figura 4.8 ilustra o conjunto *fuzzy* gerado para a variável de entrada acurácia. Definiu-se que os valores lingüísticos possíveis são $\{fraca, boa$ e $excelente\}$.

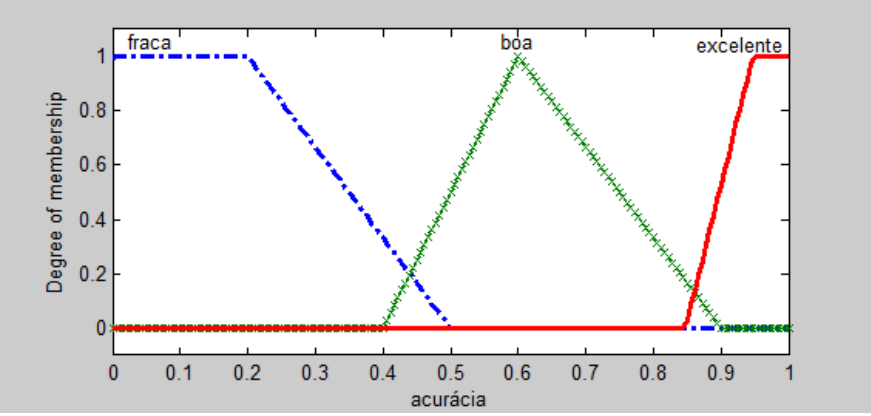


Figura 4.8: Funções de Pertinência da variável de entrada Acurácia

4.5.3. Função de fuzzificação para a variável de entrada Consistência

O mesmo se aplica para a variável de entrada consistência, também composta por três funções de pertinência (uma trapezoidal e duas triangulares). Seus conjuntos *fuzzy* apresentam a seguinte característica: limite inferior -0.1 e limite superior 1.2. A Figura 4.9 ilustra o conjunto *fuzzy* gerado para a variável de entrada consistência. Definiu-se que os valores lingüísticos possíveis são $\{baixa, média$ e $alta\}$.

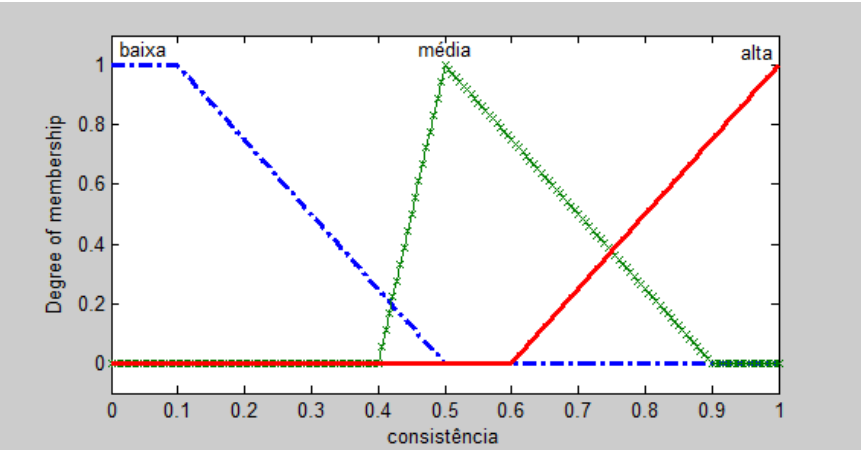


Figura 4.9: Funções de Pertinência da variável de entrada Consistência

4.5.4. Função de fuzzificação para a variável de entrada Atualidade

A Figura 4.10 ilustra a variável de entrada atualidade. Ela é composta por três funções triangulares e seus conjuntos *fuzzy* apresentam a seguinte característica: limite inferior 0 e limite superior 1.2.

Diferentemente das demais variáveis de entrada, a atualidade representa o período decorrido desde a última atualização. Conforme apresentado na Seção 4.2.3, ambientes de DW são projetados como repositório histórico da instituição. Atualizações periódicas (diária, semanal, mensal, etc.) devem ser feitas para a inclusão de novos dados. Foi considerada neste trabalho a ocorrência de atualizações semanais e, portanto, o limite superior e inferior da variável atualidade representa o número de semanas que se passaram desde a última atualização. Definiu-se que os valores lingüísticos possíveis são $\{atualizado, poucoDesatualizado$ e $muitoDesatualizado\}$.

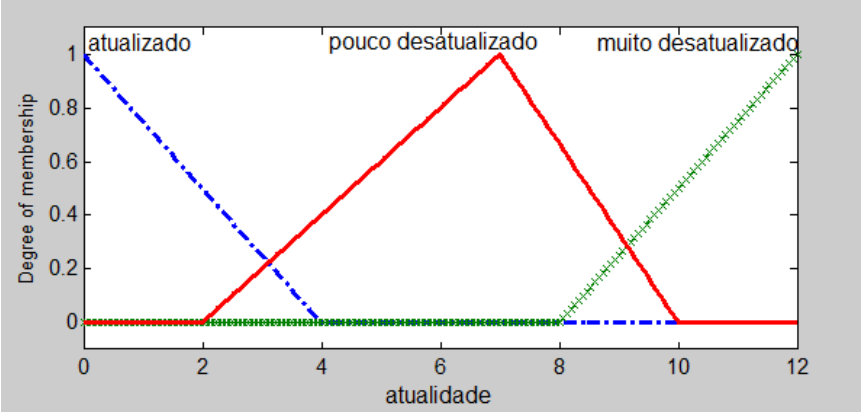


Figura 4.10: Funções de Pertinência da variável de entrada Atualidade

4.5.5. Função de fuzzificação para a variável de entrada Unicidade

Os conjuntos *fuzzy* para a variável de entrada unicidade apresentam três funções de pertinência Gaussianas e têm como limite inferior 0 e limite superior 2. A Figura 4.11 ilustra o conjunto *fuzzy* gerado para a variável. Definiu-se que os valores lingüísticos possíveis são {*único*, *duplicado* e *duplicadoInconsistente*}.

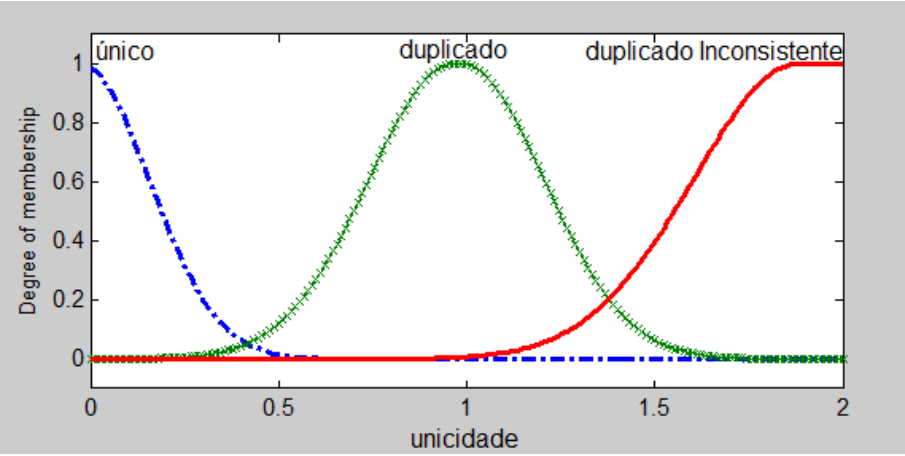


Figura 4.11: Funções de Pertinência da variável de entrada Unicidade

4.5.6. Regras de inferência para a confiança dos dados

O conjunto de regras para inferência *fuzzy*, necessário para operação do modelo, foi extraído da opinião de especialistas da área, substituindo assim, conforme já citado, as fórmulas matemáticas normalmente utilizadas. A Tabela 4.10 relaciona o conjunto de oitos regras aplicada.

Tabela 4.10: Conjunto de regras de inferência fuzzy aplicadas

No.	Entrada					Saída
	Acurácia	Completude	Unicidade	Atualidade	Consistência	Confiabilidade
#1	fraca	-	-	-	-	lixo
#2	fraca	fraca	duplicadoInconsistente	muitoDesatualizado	baixa	lixo
#3	fraca	fraca	duplicadoInconsistente	poucoDesatualizado	baixa	lixo
#4	fraca	boa	duplicadoInconsistente	poucoDesatualizado	média	baixa
#5	fraca	boa	duplicadoInconsistente	poucoDesatualizado	média	média
#6	boa	boa	duplicado	atualizado	media	boa
#7	excelente	excelente	duplicado	poucoDesatualizado	Media	excelente
#8	excelente	boa	-	atualizado	media	excelente

O conjunto de regras *fuzzy* apresentado na Tabela 4.10 representa o comportamento do sistema de confiança dos dados e opera analisando o conjunto *fuzzy* de entrada e seus impactos na avaliação da confiabilidade. Nossa intenção é demonstrar que mesmo com pequenos conjuntos de regras, i.e, apenas oito, é possível obter inferências satisfatórias.

O funcionamento de cada uma das oito regras de inferência, propostas na Tabela 4.10, pode ser visualizado na Figura 4.12.

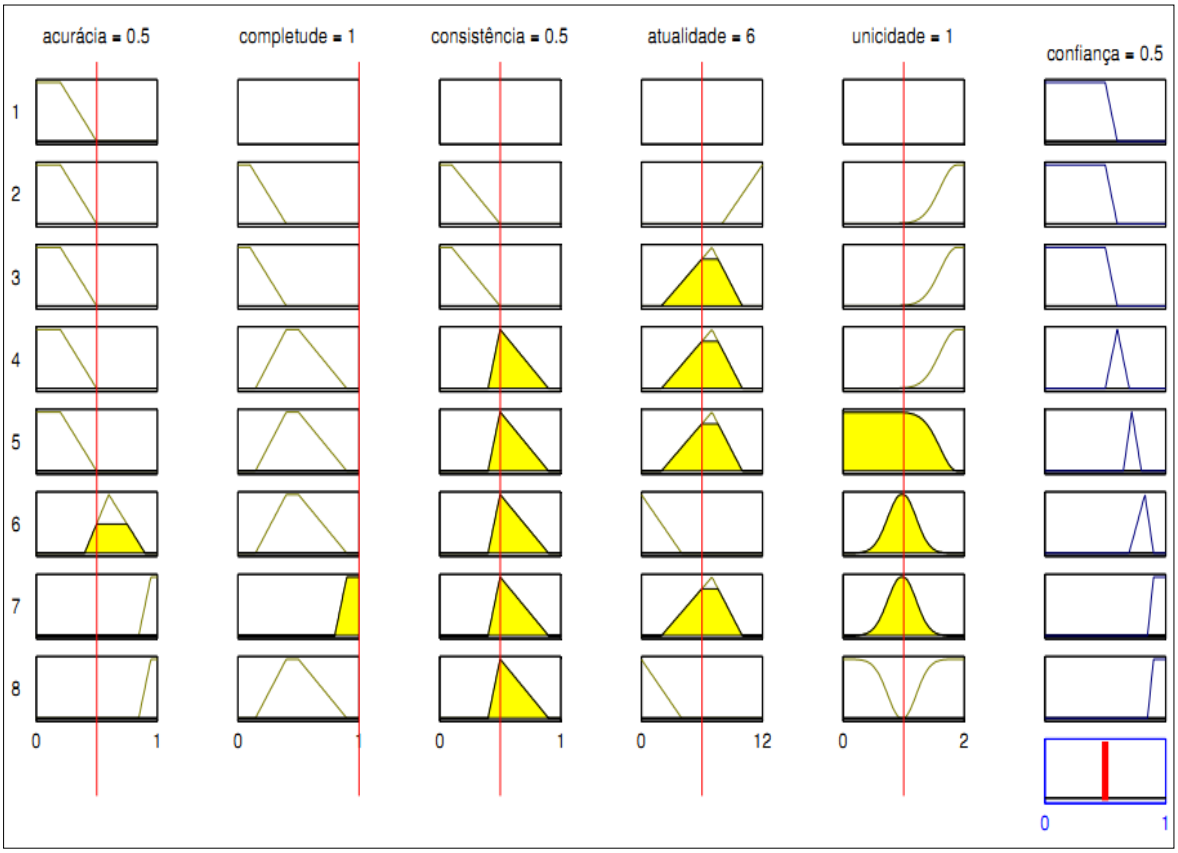


Figura 4.12: Visualização das regras do controlador fuzzy

4.5.7. Função de defuzzificação para a variável de saída confiabilidade

Como item final do modelo *fuzzy*, tem-se a etapa de *defuzzificação*. A partir da variável lingüística de controle se obtém o valor para qual o controlador executará a ação de controle desejada. No caso deste trabalho, consiste em atribuir o nível de confiabilidade para a informação de desejada.

Os conjuntos *fuzzy* adotados para a variável de saída são representados por duas funções trapezoidais e três triangulares. Elas apresentam como $limite\ inferior = 0$ e $limite\ superior = 1$.

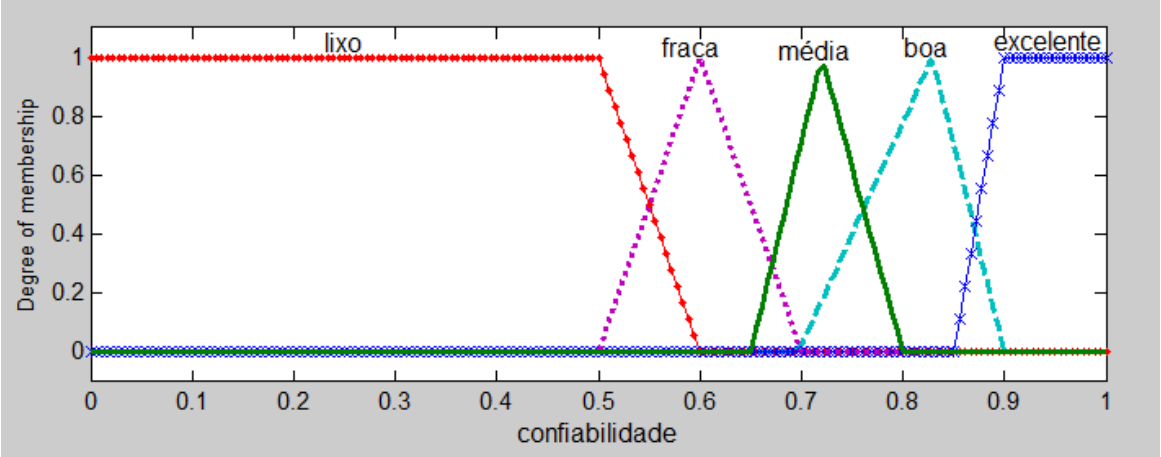


Figura 4.13: Funções de pertinência da variável de saída Confiabilidade

A Figura 4.13 ilustra o conjunto *fuzzy* gerado para a variável de saída. Definiu-se que os valores lingüísticos possíveis para representar a incerteza, através da confiabilidade são $\{lixo, fraca, média, boa$ e $excelente\}$.

Valores inferiores a 0.6 são considerados sem valor algum, e correspondem a variável lingüística $lixo$. Valores entre 0.5 e 0.7 são classificados como de $fraca$ qualidade. O intervalo correspondente a 0.65 e 0.8 caracteriza-se como de $média$ qualidade. Boa qualidade compreende o intervalo de 0.7 a 0.9 e, finalmente, $excelente$ qualidade equivale a dados com valor a partir de 0.85.

5. SIMULAÇÕES E RESULTADOS – ESTUDO DE CASO

A solução proposta será avaliada em duas etapas. Primeiramente, utilizando-se o *software* MATLAB®, em conjunto com a biblioteca *Fuzzy Logical Toolbox*™, foi simulada a entrada de alguns valores pseudo-aleatórios.

Em seguida, os resultados obtidos permitiram a realização da segunda etapa, que consistiu na incorporação do modelo *fuzzy* de confiabilidade de dados, proposto no Capítulo 4, ao ambiente de *Business Intelligence* da SPU/MP (Secretaria do Patrimônio da União; do Ministério do Planejamento, Orçamento e Gestão). Para esta implementação, utilizou-se o *software* de BI *open source Suite Pentaho* [21]. Mais adiante serão apresentadas as justificativas para esta escolha ferramental.

5.1. SIMULAÇÕES E RESULTADOS

A Tabela 5.1 apresenta um conjunto de dez experimentos realizados através da entrada de alguns valores pseudo-aleatórios para cada uma das cinco dimensões de qualidade identificadas (conforme Seção 4.5). Na última coluna os resultados obtidos para a variável de saída *Confiabilidade* são apresentados.

Tabela 5.1: Resultado dos Experimentos Realizados

No. do experimento	Entrada					Saída
	D_{Comp}	D_{Acur}	D_{Cons}	D_{Atua}	D_{Unic}	$Confiabilidade$
#1	2	1	1	1	1	36.95
#2	6.3	2	10	2.2	1	33.90
#3	9.6	3.5	10	6	1	45.92
#4	4.5	7	2.5	2	0	80
#5	4	2	5	5	20	46.27
#6	10	2	10	5	20	38.28
#7	0	0	0	12	0	27.60
#8	5	5	5	5	15	87.58
#9	1	1	1	1	1	35.24
#10	10	10	10	10	10	90.25

D_{Comp} = Dimensão da Completude
D_{Acur} = Dimensão da Acurácia
D_{Cons} = Dimensão da Consistência
D_{Atua} = Dimensão da Atualidade
D_{Unic} = Dimensão da Unicidade

5.1.1. Análise e Discussão dos Resultados

Como pode ser observado na Tabela 5.1, algumas dimensões apresentaram maior nível de influência sobre a confiabilidade dos dados. Isso é facilmente observado nos experimentos #2, #5 e #6, nos quais a baixa pontuação para a Acurácia (ou seja, valor da Acurácia=2 e Acurácia=3.5) transformou o resultado para a Confiabilidade em valores inferiores a 60.0 (i.e., equivalente a variável lingüística de saída $Lixo$).

Assim, conforme expresso nos resultados apresentados, o método proposto demonstrou coerência em relação às regras de inferência para a aplicação hipotética proposta neste trabalho. Desse modo, a tarefa de medição dos níveis de confiança dos dados, a partir de medidas de completude, acurácia, consistência, atualidade e unicidade foi bem sucedido.

As Figuras 5.1 e 5.2 mostram as superfícies obtidas a partir das regras para a completude, acurácia, consistência, unicidade, atualidade e a variável de saída confiabilidade. Cada figura mostra a associação entre duas variáveis lingüísticas. O resultado final corresponde à confiabilidade. Baseado nessas figuras é possível analisar a contribuição de cada variável lingüística para a confiabilidade dos dados e, se necessário, as regras podem ser ajustadas para representar melhor o contexto.

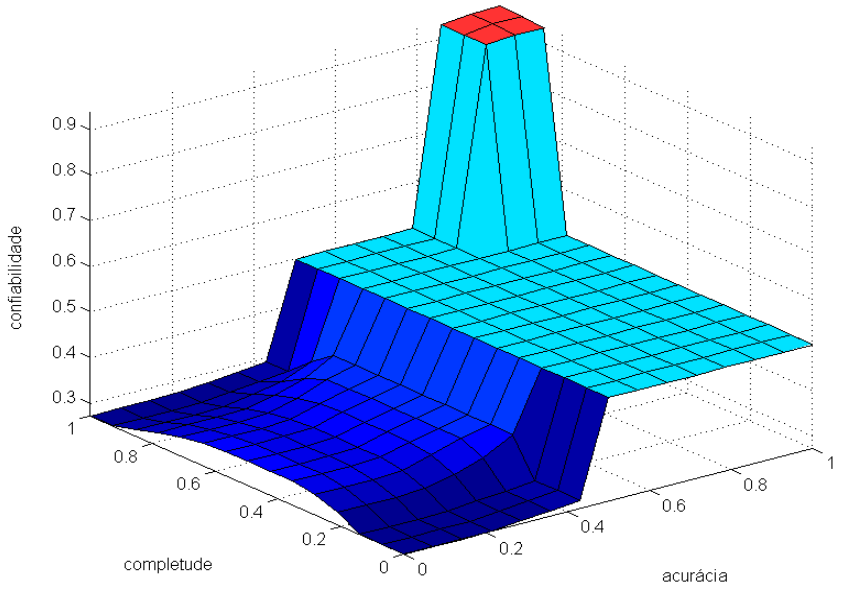


Figura 5.1: Análise de sensibilidade da completude e da acurácia em relação a confiabilidade

Através da Figura 5.1 pode-se observar a relação entre a dimensão da completude e da acurácia. Como discutido na seção 4.5.6, o impacto da acurácia é maior sobre a confiabilidade dos dados.

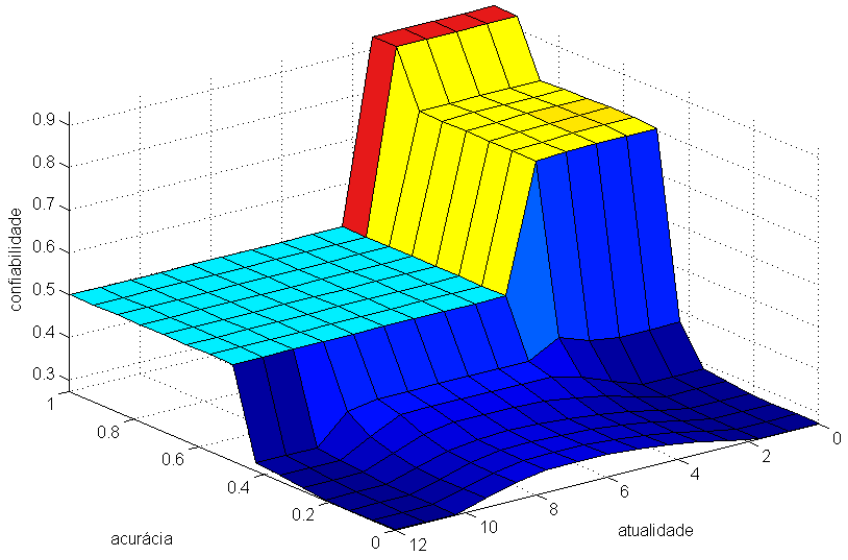


Figura 5.2: Análise de sensibilidade da acurácia e da atualidade em relação a confiabilidade

A Figura 5.2 ilustra uma segunda relação entre as dimensões da qualidade, através da comparação entre a acurácia e da atualidade. Apesar de a acurácia ser decisiva na confiabilidade dos dados, é possível observar que dados atualizados precisam ser considerados de extrema importância para a QD.

5.2. EXEMPLO DE APLICAÇÃO: CONFIABILIDADE DOS DADOS NA SECRETARIA DO PATRIMÔNIO DA UNIÃO (SPU)

Nesta seção será apresentado um exemplo de aplicação, que consiste na incorporação do Modelo de Confiabilidade de Dados, proposto no Capítulo 4, ao ambiente de *Business Intelligence* da Secretaria de Patrimônio da União, do Ministério do Planejamento, Orçamento e Gestão (BI-SPU), para fornecer informações sobre a confiabilidade dos dados.

5.2.1. Descrição do Cenário

O cenário de exemplo de aplicação é a Secretaria de Patrimônio da União, do Ministério do Planejamento, Orçamento e Gestão (SPU/MP). A SPU é um órgão da administração direta vinculado ao Ministério do Planejamento, Orçamento e Gestão (MP), e tem por missão "conhecer, zelar e garantir que cada imóvel da União cumpra sua função sócio-ambiental em harmonia com a função arrecadadora, em apoio aos programas estratégicos para a Nação".

O patrimônio da União, de natureza tão diversificada, encontra-se atualmente classificado em imóveis próprios nacionais, terrenos de marinha, áreas de preservação permanente, terras indígenas, florestas nacionais, terras devolutas, áreas de fronteira e bens de uso comum.

Atualmente, a SPU dedica-se em gerir um patrimônio de mais de 600 mil imóveis, localizados em toda a extensão territorial da Federação. Nesse sentido, a SPU desempenha papel fundamental, disponibilizando áreas vazias ou subutilizadas da União para o desenvolvimento de projetos de provisão de moradia para a população. É nesse contexto que será inserido o exemplo de aplicação apresentado neste capítulo. Este estudo de caso permitirá avaliar a aplicabilidade do modelo de confiabilidade proposto no Capítulo 4.

5.2.2. Arquitetura do ambiente de BI da SPU

Os dados primários (OLTP), a partir dos quais se compõem as informações gerenciais disponibilizadas no ambiente de BI-SPU, estão fragmentados em diversas fontes de dados, estabelecidas sob plataformas de *hardware* e *software* diferentes, apresentando problemas específicos de QD. A relação completa das fontes de dados utilizados pelo BI-SPU é apresentada a seguir:

- **SIAPA:** Sistema Integrado de Administração Patrimonial. É o sistema onde são cadastrados os imóveis dominiais da União, suas utilizações e os eventos financeiros;
- **SPIUNET:** O Sistema de Gerenciamento do Patrimônio Imobiliário de uso especial da União faz a gerência da utilização dos imóveis da União, de caráter "Bens de Uso Especial". Imóveis pertencentes a esta categoria, são pertencentes à União (Administração Pública Federal Direta), de terceiros que a União

utiliza, próprios de Fundações e Autarquias e de Empresas Estatais dependentes;

- **Ambiente de DW legado:** Sistema de DW legado da SPU que é alimentado pelos sistemas principais, SIAPA e SPIUNET. Atualmente tem sido utilizado como sistema emissor de relatórios gerenciais;
- **Planilhas e relatórios gerenciais:** São planilhas eletrônicas e/ou relatórios gerenciais desenvolvidos e disponibilizados por alguma Gerência Regional do Patrimônio da União (GRPU). Cada Estado da Federação dispõe de uma GRPU;
- **Indicadores sociais do IBGE:** O Instituto Brasileiro de Geografia e Estatística (IBGE) fornece informações referentes à população, sociedade, economia e território da União.

A Figura 5.3 ilustra com maiores detalhes a arquitetura metodológica de carga adotada no ambiente de *Business Intelligence* da SPU (BI-SPU).

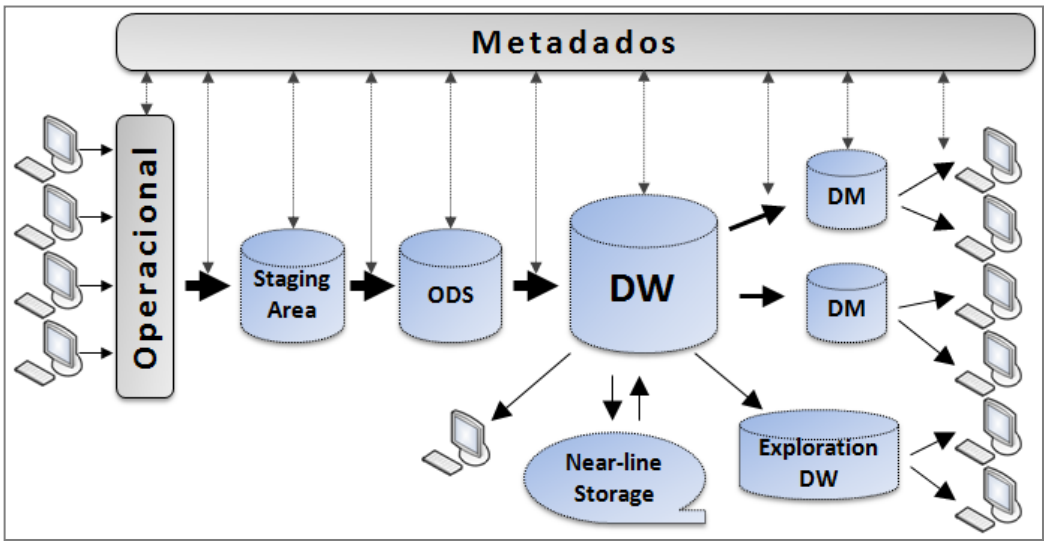


Figura 5.3: Arquitetura metodológica de carga no BI da SPU

5.2.2.1. Ferramentas Open Source de BI - A Suite Pentaho

A *Open Source Initiative* (OSI) é uma corporação sem fins lucrativos dedicada a administrar e promover a definição de *softwares open source* (ou código aberto). Para ser considerado um software *open source*, a OSI exige a disponibilização do código fonte dos programas, segundo os critérios estabelecidos na *Open Source Definition* [20].

Dentre as soluções *Open Source* de BI existentes, como SpagoBI, JasperSoft e BIRT [84], o projeto Pentaho BI é considerado uma das melhores ferramentas na área que se tem conhecimento, chegando a superar, inclusive, projetos proprietários de grandes empresas [21, 22]. Sua estrutura modular baseada na linguagem Java[3] permite o desenvolvimento de novas extensões de *software* interoperáveis.

Por meio do repositório de projetos *open source* SourceForge.net, um dos maiores da Internet, é possível constatar um percentual de atividade da ferramenta superior a 99% [23]. A Figura 5.4 apresenta uma representação da arquitetura da Suite Pentaho BI. A Suite Pentaho é uma solução completa de BI e é composta das seguintes ferramentas [21]:

- **Pentaho Data Integration:** Conhecida também como Kettle, é a solução para integração de dados, recomendada para processos de ETL (do inglês *Extract, Transfomation and Load*) responsável por popular o *Data Warehouse*, Migração de base de dados e Integração entre Aplicações
- **Pentaho Analysis:** Também conhecida como Mondrian, é um motor OLAP, baseado em uma arquitetura ROLAP (do inglês *Relational Online Analytical Processing*). Possui funcionalidades, como, camada de metadados, linguagem MDX, *cache* em memória, tabelas agregadas entre outros.
- **Pentaho Reporting:** Ferramenta responsável pela geração de relatórios. É derivado do projeto *JFreeReport* e é apoiado por outras ferramentas para a geração de metadados, a qual permite a criação *ad hoc* de relatórios, via navegador *web*.
- **Pentaho Dashboards:** Permite a criação de painéis de controle, mais conhecidos como *Dashboards*. Através dele é possível reunir, em uma mesma tela, os principais indicadores.
- **Pentaho Data Mining:** Também conhecido como Weka Data Mining, é o módulo para recursos de mineração de dados.

[3] Java é uma Linguagem de Programação Orientada a Objetos. Disponível em <http://www.java.com>.

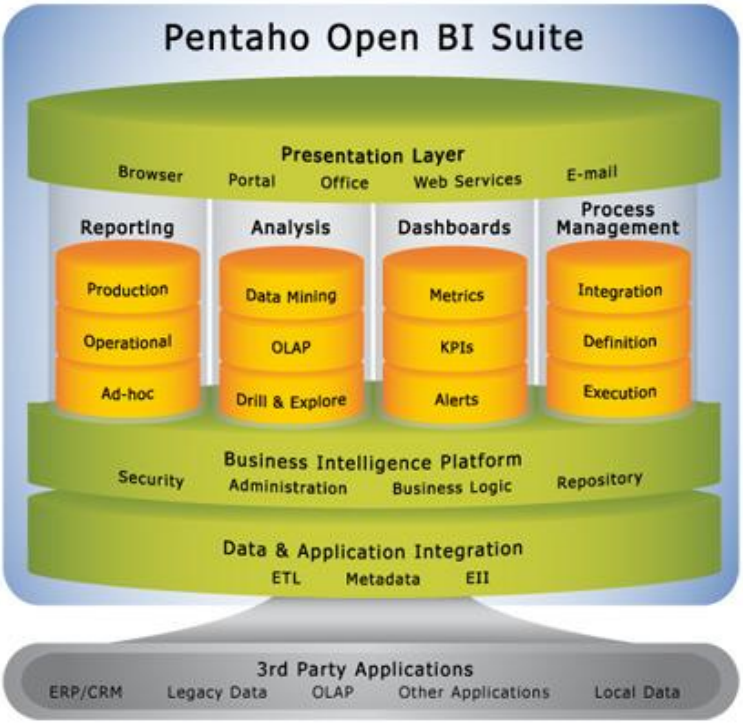


Figura 5.4: Arquitetura da Suite Pentaho BI. Obtida de (Pentaho)

A Pentaho Corporation [21] é a principal patrocinadora do projeto Pentaho BI. Este projeto *open source* é uma iniciativa que provê suporte às necessidades das empresas através de ferramentas de suporte à decisão, e abrange as seguintes áreas: Relatórios, Análises, *Dashboards*, Mineração de Dados, *Workflow* e Servidor Plataforma de BI (Servidor), conforme Figura 5.4.

5.2.2.2. Modelo Multidimensional

O BI-SPU é composto de diversos modelos multidimensionais, haja vista a complexidade de variáveis presentes na gestão dos imóveis da União. Entre os modelos multidimensionais existentes, fez-se necessário a escolha de um cubo para avaliação da nossa solução. O cubo escolhido foi o de Desempenho Organizacional, apresentado na Figura 5.5.

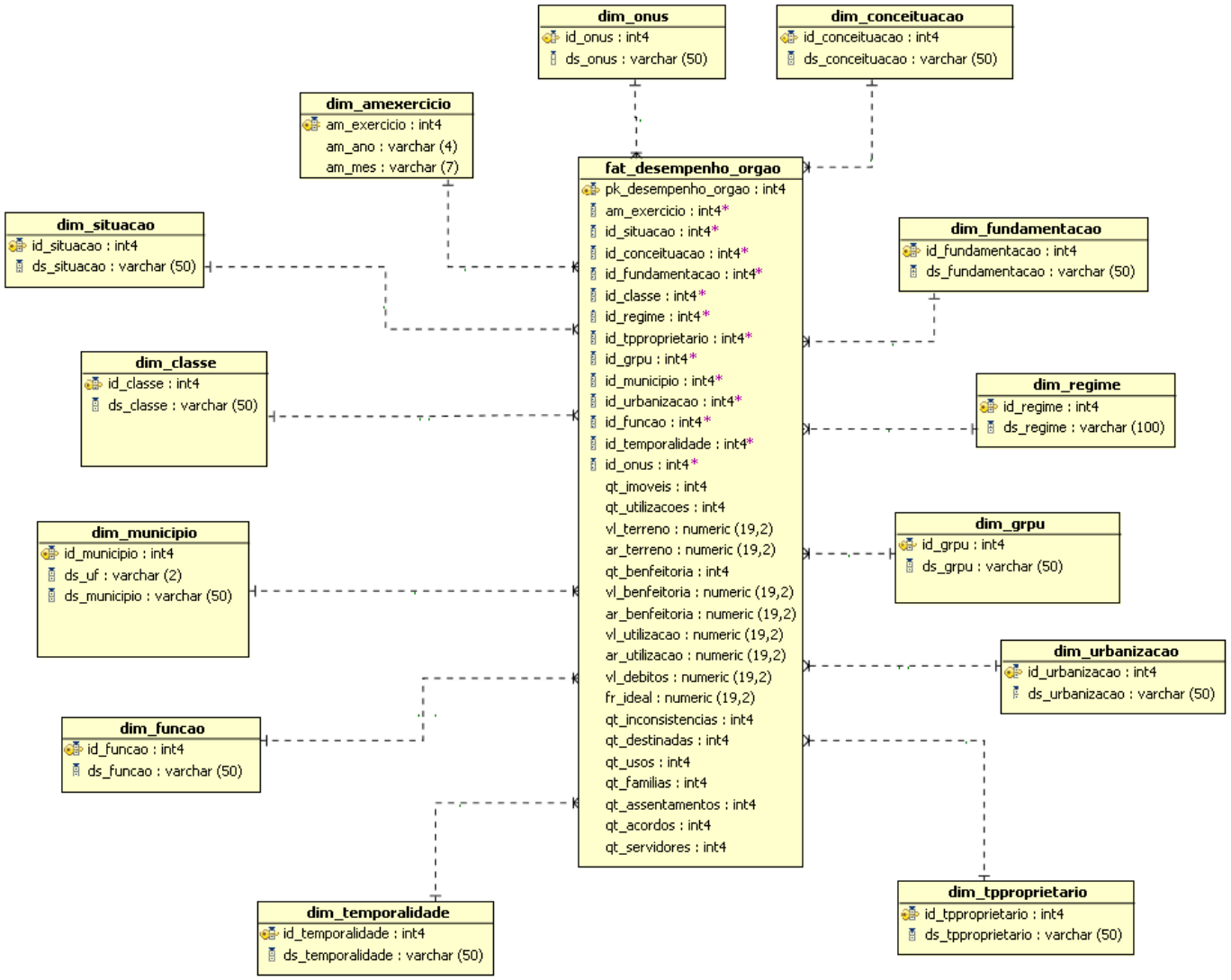


Figura 5.5: Modelo Multidimensional da SPU/MP – Cubo Desempenho Organizacional

O cubo Desempenho Organizacional tem por objetivo avaliar o desempenho organizacional do Órgão Central da SPU (localizado no Distrito Federal) e de suas 27 Superintendências Regionais (conhecidas como GRPU) espalhadas pelas Unidades da Federação. Este cubo é composto de 13 dimensões: Temporalidade, Função, Município, Classe, Situação, Ano e Mês de Exercício, Ônus, Conceituação, Fundamentação, Regime, GRPU, Urbanização e Tipo de Proprietário.

5.2.3. Metadados de Confiabilidade dos dados

Os dados encontram-se armazenados em diferentes níveis de granularidade. A definição de métricas por si só não assegura avaliações precisas dos dados. É necessário analisar os dados em níveis de granularidade adequados. Admitindo que não existam limitações para armazenamento dos dados, cada campo *c'* do banco de dados receberia os metadados nos formatos definidos na seção 4.3, conforme Requisitos das métricas de QD.

Para realizar o processo de descoberta dos metadados de qualidade, será utilizado o método conhecido como Perfil dos dados (do termo inglês *Data Profiling*). O processo de *Data Profiling* pretende detectar de forma sistemática os erros, as inconsistências, as redundâncias e a existência de informação incompleta nos dados [62, 63]. A partir da taxonomia proposta no Capítulo 4, a detecção dos problemas produzirá um conjunto de metadados que irá compor cada uma das cinco dimensões da confiabilidade dos dados.

5.2.3.1. Armazenamento dos metadados

O método de armazenamento dos metadados de confiabilidade dos dados traz impacto direto na performance do DW. Na literatura, duas soluções têm sido propostas para tratar dessa importante questão [41]. A primeira proposta sugere a materialização dos metadados em repositórios, enquanto a segunda abordagem defende a não materialização, também citado por alguns autores como geração de metadados em tempo real.

Amaral (2003) [41] esclarece que 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 probabilidade de ocorrerem inconsistências, que poderiam ser eliminadas na etapa de ETL do DW. Em contrapartida, a não materialização do esquema diminui o intervalo de tempo necessário

para tornar os dados disponíveis para consulta, porque elimina uma etapa do processo de ETL, na carga dos dados no DW.

O impacto no desempenho do DW em relação solução escolhida reside, portanto, na relação custo de processamento *versus* espaço de armazenamento. Enquanto a abordagem virtual proporciona uma grande economia de espaço em disco, torna crítico, por outro lado, a questão do processamento em tempo real. Essa tem sido uma das grandes dificuldades para estruturas de metadados.

Neste trabalho, optou-se pela materialização dos metadados, através da criação tabelas auxiliares para o armazenamento dos metadados. A geração de metadados em tempo real é uma importante linha de pesquisa em atividade na atualidade e sua adoção em nosso modelo será apresentada na seção 6.1, como próximos passos nos trabalhos futuros de nossa pesquisa.

A Figura 5.6 apresenta nosso modelo de armazenamento de metadados de qualidade em tabelas auxiliares. A solução é inspirada por Boyadzhieva et al. [64] e oferece a separação física entre os dados e os metadados. Sua grande vantagem reside no fato de oferecer maior flexibilidade para a manipulação de ambos os repositórios (em contraste com o método de extensão das tabelas do modelo físico e lógico do DW).

FK_tabela_fato	completude	atualidade	consistencia	unicidade	acuracia

Figura 5.6: Modelo de Tabela Auxiliar para metadados de Confiabilidade dos Dados

A medida que o tomador de decisão, utilizando o módulo de consulta OLAP do BI-SPU executa uma consulta, ele tem a possibilidade de consultar também informações sobre a confiabilidade dos dados. Deste ponto em diante, o processo *fuzzy* de avaliação da confiabilidade dos dados é executada, conforme descrito na seção 4.5.

A simplicidade de sua implementação, no entanto, não a torna limitada em relação à capacidade de operação junto aos mecanismos de análise de dados. Para ilustrar a manipulação de consultas OLAP, apresentamos na Figura 5.7 um modelo de cubo de confiabilidade dos dados, em relação a duas dimensões, i.e, temporalidade e localização geográfica (i.e., Dimensão Região).

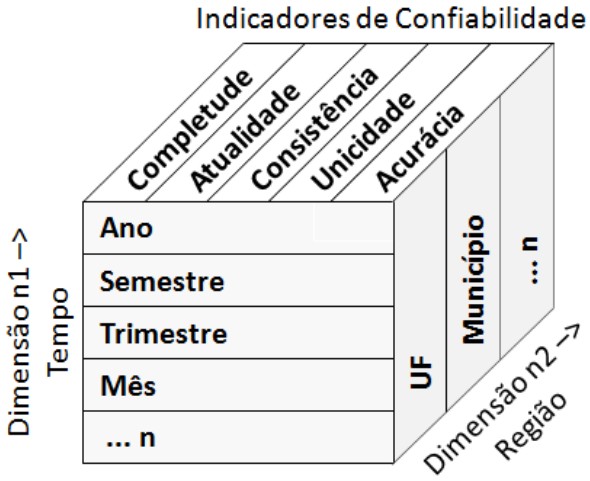


Figura 5.7: Cubo de Confiabilidade dos Dados

5.2.4. Consulta a Confiabilidade dos Dados no ambiente da BI-SPU

Administrar o patrimônio imobiliário da União é uma das grandes preocupações do Governo Federal, representada em nosso governo pela Secretaria de Patrimônio da União (SPU).

Por intermédio do portal BI-SPU, os gestores dispõem de informações a respeito do patrimônio imobiliário da União, utilizadas para a tomada de decisão, tais como: quantidade de imóveis destinados, quantidade de imóveis demarcados e homologados, quantidade de imóveis adquiridos por dação em pagamento, quantidade de imóveis dominiais cadastrados do município, entre outros. Para maior segurança em relação à decisão que será tomada, é fundamental que o resultado das consultas possa ter sua qualidade confrontada por intermédio de informações sobre a confiabilidade dos dados disponibilizados.

A seguir, é apresentado um exemplo de interação junto ao Portal de *Business Intelligence* da SPU e a obtenção de informações de confiabilidade sobre uma consulta realizada a respeito do Desempenho Organizacional:

"SELECIONAR
A QUANTIDADE DE DESTINAÇÕES, A ÁREA DE UTILIZAÇÃO E
O VALOR DE UTILIZAÇÃO DO IMÓVEL
OCORRIDA NOS MESES DE
OUTUBRO, NOVEMBRO E DEZEMBRO DE 2011".

Esta consulta é expressa na sintaxe da linguagem para consultas multidimensionais MDX[4] da seguinte forma:

```
SELECT NON EMPTY
     {[Measures].[Qt_Destinacoes],
      [Measures].[Area_Utilizacao],
      [Measures].[Valor_Utilizacao]} ON COLUMNS,
NON EMPTY
     {[Tempo].[All_Tempo].[2011].[10/2011],
      [Tempo].[All_Tempo].[2011].[11/2011],
      [Tempo].[All_Tempo].[2011].[12/2011]} ON ROWS
FROM [fatdesemporg]
```

Onde:

`fatdesemporg` é o cubo de Desempenho Organizacional do DW
`Qt_Destinacoes` são variáveis de fato do cubo `fatdesemporg`
`Area_Utilizacao`
`Valor_Utilizacao`
`Tempo` é a dimensão da Temporalidade do cubo `fatdesemporg`

Considerando que a consulta em MDX seja submetida via módulo de consulta OLAP, além do retorno da resposta solicitada, será possível obter o nível de confiabilidade dos dados conforme Tabela 5.2, obtido através da arquitetura proposta nesta dissertação. Destaca-se que o formato de apresentação da confiabilidade para o usuário final será definido pela ferramenta cliente, que neste exemplo é o Servidor OLAP da Suite Pentaho BI, conhecida como Mondrian [21].

Tabela 5.2: Confiabilidade dos fatos em relação a dimensão Tempo

Fato / Dimensão Tempo	Qt_Destinacoes	Area_Utilizacao (m^2)	Valor_Utilizacao (R$)	Confiabilidade
10/2011	511	12.776.330,02	30.505.237,50	88,65
11/2011	333	2.573.438.715,01	444.090.196,95	84,50
12/2011	252	1.764.940,73	41.476.669,38	89,00

[4] MDX é uma linguagem de consulta a objetos multidimensionais (como cubos) que retornam conjuntos de células multidimensionais, contendo dados do cubo. Tornou-se a linguagem padrão de aplicações analíticas.

5.2.5. Considerações sobre a consulta da Confiabilidade

Sobre a consulta exemplificada na seção 5.2.4, pode-se concluir que tanto a dimensão $Tempo$, quanto as variáveis de fatos $Qt\_Destinacoes$, $Area\_Utilizacao$ e $Valor\_Utilizacao$ contêm informações bastante confiáveis, conforme evidenciado pelos graus de confiabilidade desses objetos. Mesmo para o resultado com menor confiabilidade, i.e., mês $11/2011$, é possível tomar decisões baseadas em informações confiáveis por haver obtido um valor de confiabilidade igual a $84{,}50$ (valor lingüístico para confiabilidade = Boa, conforme seção 4.5.7).

Vale ressaltar que mesmo com as limitações, este exemplo de aplicação foi capaz de demonstrar a contribuição de desta proposta.

6. CONCLUSÕES

Este trabalho apresentou um conjunto de técnicas para avaliação dos dados em ambientes de *Business Intelligence*. O objetivo foi a definição de um modelo *fuzzy* de avaliação da confiabilidade dos dados a partir da investigação dos problemas de qualidade manifestos em repositórios de dados. Inicialmente, foram apresentadas as definições para BI. Em seguida, uma revisão bibliográfica detalhada sobre a questão da qualidade e confiabilidade dos dados foi apresentada no Capítulo 3.

A partir do Capítulo 4, a proposta foi apresentada e estruturada em quatro etapas, objetivando a apresentação de três contribuições. A primeira contribuição foi apresentada logo em seguida. A partir de uma análise cuidadosa, uma proposta de taxonomia para os problemas de QD em modelos multidimensionais foi desenvolvida e apresentada, compondo-se de cinco dimensões de qualidade.

A segunda contribuição foi a definição de métricas de QD abrangentes e objetivas, sempre que possível, para avaliação da qualidade dos dados. Definições e exemplos foram apresentados para evitar diferentes interpretações. Por fim, apresentamos um novo método de avaliar a confiabilidade e a noção da incerteza dos dados, através da agregação das pontuações de qualidade apoiada pela lógica *fuzzy*. Foi demonstrado, através de experimentos, que esta abordagem é satisfatória para a avaliação da confiabilidade dos dados em ambientes de *Business Intelligence* e *Data Warehouse*, por meio de simulações e um exemplo de aplicação, junto a Secretaria de Patrimônio da União (SPU) mostrando-se útil e contribuindo para a evolução dos processos de gestão do patrimônio da União.

6.1. SUGESTÕES PARA TRABALHOS FUTUROS

A área de pesquisa em qualidade e confiabilidade de dados é muito vasta, tanto no que se refere ao desenvolvimento e à aplicação de técnicas de gerenciamento dos dados, como no que se refere à avaliação da qualidade.

Conforme discutido na Seção 1.2, o trabalho desenvolvido e apresentado nesta dissertação não tem a pretensão de fornecer soluções definitivas, mas apenas contribuir com uma pequena parcela de conhecimento, a fim de permitir algum avanço técnico-científico na área de estudo em questão. Considerando a perspectiva da avaliação da confiabilidade, este trabalho apresenta algumas limitações e deixa alguns problemas em

aberto, que poderão ser resolvidos em trabalhos futuros. A seguir são apresentadas algumas sugestões:

- Os resultados obtidos na avaliação da confiabilidade dos dados poderiam ser utilizados para influenciar a QD, resultando assim, na imposição de mudanças nos dados e nos componentes;
- Incorporação do modelo de confiabilidade dos dados em algum tipo de ciclo de vida da QD. Modelos de gerenciamento como o TDQM, por exemplo, poderiam ser beneficiados pela medição da confiança, em seu ciclo para a melhoria da qualidade da informação;
- Definição de taxonomias e indicadores de confiabilidade para outros objetos da arquitetura de BI, além dos valores dos dados no modelo multidimensional, e a especificação de componentes para a medição da qualidade dos mesmos;
- Os metadados de QD foram obtidos em nossa proposta através de sua materialização em tabelas auxiliares. A geração de metadados em tempo real é uma importante linha de pesquisa em atividade na atualidade e sua adoção poderia em nosso modelo traria alguns benefícios, conforme Seção 5.2.3.1.
- Em relação à avaliação experimental (i.e., estudo de caso), a implementação das funcionalidades necessárias para a interação com a plataforma de BI em uma ferramenta de consulta analítica. Uma ferramenta de "Confiança BI" (do termo em inglês, *TRUST-BI*) habilitaria o sistema receber as informações de qualidade, e apresentá-las para o usuário final juntamente com os resultados da consulta;

Além disso, já está em andamento a implementação de um protótipo do modelo de confiabilidade dos dados para integrar-se com a suite *open source* Pentaho BI.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Sadiq, S.; Yeganeh, N. K.; and Indulska, M. "20 years of data quality research: Themes, trends and synergies," In *Proceedings of the 22nd Australasian Database Conference (ADC 2011)*. Sidney, Austrália, pp. 1-10, 2011.

[2] Daniel, F.; Casati, F.; Palpanas, Th.; Chayka, O. "Managing Data Quality in Business Intelligence Applications," In *Proc. of the 6th International Workshop on Quality in Databases (QDB 08) at VLDB*, pp. 133-143, 2008.

[3] Kashyap, V. "Trust, But Verify: Emergence, Trust, an Quality in Intelligent Systems," *IEEE Intelligent Systems*, vol. 19, no. 5, pp. 85-87, 2004.

[4] Boll, S.; Cappiello C.; Gertz, M.; Kashyap, V.; Sattler, K.; and Zeigler, C. "Do you Trust in Data Quality?," *Working Group on Trust and Data Quality, Dagstuhl Workshop on Data Quality*, nov. 2003. Disponível em: <http://sirius.cs.ucdavis.edu/Dagstuhl03/wgs/03362.KashyapVipul.Other.ppt>. Acesso em: 10 jan. 2010.

[5] Gertz, M.; Ozsu, M. T.; Saake, G.; and Sattler, K.-U. "Report on the Dagstuhl Seminar 'Data Quality on the Web' ". *SIGMOD RECORD*, vol. 33, no. 1, pp. 127-132, 2004.

[6] Rodriguez, C.; Daniel, F.; Casati, F.; Cappiello, C. "Toward Uncertain Business Intelligence: The Case of Key Indicators". *Journal of IEEE Internet Computing*, vol. 14, no. 4, pp. 32-40, 2010.

[7] Diane M. Strong, Yang W. Lee, Richard Y. "Wang: Data Quality in Context". *CACM*, 40(5): 103-110, 1997.

[8] Wang, R. Y.; Strong, D. M.; "Beyond Accuracy: What Data Quality Means to Data Consumers?", *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5-34, 1996.

[9] Kahn, B. K.; Strong, D. M.; Wang, R. Y. "Information Quality Benchmarks: Product and Service Performance". *Communications of the ACM*, vol. 45, no. 4, pp. 184-192, 2002.

[10] Wand, Y.; and Wang, R. Y. "Anchoring Data Quality Dimensions Ontological Foundations," *Communications of the ACM*, vol. 39, no. 11, pp. 86-95, 1996.

[11] Angel, H. "Identifying and Prioritizing Information Quality Dimensions for Assurance in the Pre-Processing Stage of Data Storage for Business Intelligence". Master's thesis, University of Oregon, Oregon, 2011.

[12] Negash. S. “Handbook on decision support systems 1: Business intelligence”. In *International handbooks on information systems*. (chapter 45). Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-540-48713-5, 2008.

[13] Yeganeh, N.; and Sadiq, S. “Avoiding Inconsistency in User Preferences for Data Quality Aware Queries,” *Business Information Systems*, 2010.

[14] Richardson, M.; Agrawal, R.; and Domingos, P. “Trust Management for the Semantic Web,” *Proceedings of the 2nd International Semantic Web Conference (ISWC)*, Sanibel Island, Florida, October 2003.

[15] Chaudhuri, S.; Dayal, U. “An Overview of Data Warehousing and OLAP technology,” ACM *SIGMOD Record*, pp. 65-74, 1997.

[16] Gartner Inc. “Using Business Intelligence to Gain a Competitive Edge,” *A special report.* Gartner, Inc.: Stamford CT, 2004

[17] Hussain, F.; Chang, E.; and Dillon, T. S. “Trustworthiness and CCCI Metrics for Assigning Trustworthiness in P2P Communication,” *International Journal of Computer Systems Science and Engineering*, vol. 19, no 4, pp. 95-112, 2004.

[18] Montaño, R. A. N. R.; Miranda, M. R.. Castilho, M. A.; Silva, L. F.; Hexsel, R. A. “Businness Intelligence nas Escolas Públicas do Estado do Paraná”, *IX Workshop de Software Livre (WSL'2008), in Anais do nono Workshop de Software Livre*, Porto Alegre, RS, 2008.

[19] Sottara, D.; Mello, P.; and Proctorm M. “Configurable Rete-OO Engine for Reasoning with Different Types of Imperfect Information”. *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, 2010.

[20] Open Source Definition. Disponível em: <http://www.opensource.org/docs/definition.php>. Acesso em: 12 de dez. 2009.

[21] Pentaho Open Source Business Intelligence. Disponível em: <http://community.pentaho.com>. Acesso em: 12 de dez. 2009.

[22] Cramer, R. “Estudo Analítico de Ferramentas Open Source para Ambientes OLAP”. Monografia (Especialização em Gerenciamento de Banco de Dados), Universidade do Extremo Sul Catarinense, Criciúma-SC, 2006.

[23] SourceForge.Net. Disponível em: <http://sourceforge.net/projects/pentaho>. Acesso em: 07 de jun. de 2010.

[24] Andersson, D.; Fries, H.; and Johansson, P. “Business intelligence: The impact on decision support and decision making processes” (Unpublished Master’s thesis). Jonkoping University, Norway, 2008. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:hj:diva-1159

[25] Pawluk, P. “Trusted data in IBM’s MDM: Accuracy dimension”, *Proceedings of International Multiconference on Computer Science and Information Technology (IMCSIT-ECOM&EGOV'10),* 2010.

[26] Pawluk , P.; Gryz, J.; Hazlewood, S.; van Run, P. “Trusted Data in IBM’s Master Data Management”, *in Proceedings of the Second International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA'10),* 2011.

[27] Gamble M.; Goble C. “Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model*”. ACM International Conference on Web Science*, pp. 1-8, 2011.

[28] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” *ACM Computing Surveys*, vol. 41, no. 3, 2009.

[29] Calazans, A. T. S. Qualidade da informação: conceitos e aplicações. *Transinformação*, v. 20, n. 1, 2008.

[30] Caro, A.; Calero, C.; and Piattini, M. “Assessment of Web Portal Data Quality using Bayesian Networks”, In *33° Conferencia Latinoamericana de Informática (CLEI07)*, San José, Costa Rica, pp. 57-66, 2007.

[31] Cappiello, C.; Francalanci, C.; and Pernici, B. “Data quality assessment from the user’s perspective”. In *International Workshop on Information Quality in Information Systems, (IQIS2004)*. Paris, France: ACM, 2004.

[32] Wang R., “A Product Perspective on Total data Quality Management”. *Communications of the ACM*, vol. 41, no. 2, pp. 54-65, 1998.

[33] Lee, Y. “AIMQ: a methodology for information quality assessment. Information and Management”, *Elsevier Science*, pp. 133-146, 2002.

[34] Bouzeghoub, M.; and Kedad, Z. “Quality in Data Warehousing, in Information and Database Quality”, M. Piattini, C. Calero, and M. Genero, Editors. Kluwer Academic Publishers, 2001.

[35] Burgess, M.; Fiddian, N.; and Gray, W. “Quality Measures and the Information Consumer”. In *Proceeding of the Ninth International Conference on Information Quality*, 2004.

[36] Knight S. A.; and Burn, J. M. “Developing a Framework for Assessing Information Quality on the World Wide Web”. *Informing Science Journal*, pp. 159-172, 2005.

[37] Eppler, M. J. “Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes,” Springer, 2nd Ed., 2003.

[38] Malak, G.; Sahraoui, H.; Badri, L.; and Badri, M. “Modeling Web-Based Applications Quality: A Probabilistic Approach”. In *7th International Conference on Web Information Systems Engineering*, Wuhan, China: Springer LNCS, 2006.

[39] Naumann, F. “Quality-driven query answering for integrated information systems,” Springer-Verlag New York, Inc., New York, NY, USA, 2002.

[40] Pipino, L.; Wang, B.; Lee, Y. “Data Quality Assessment”. *Communications of the ACM*, vol. 45, no. 4, pp. 211-218, 2002.

[41] Amaral, G. C. M. “Aquaware: Um Ambiente de Suporte à Qualidade de Dados no Data Warehouse”. Dissertação (Mestrado em Informática), Universidade Federal do Rio de Janeiro - UFRJ, Rio de Janeiro-RJ, 2003.

[42] Vassiliadis, P.; Bouzeghoub, M.; and Quix, C. “Towards Quality-Oriented Data Warehouse Usage and Evolution”. *Proceedings 11th Conference of Advanced Information Systems Engineering (CAiSE ’99)*, Heidelberg, Germany. 1999.

[43] Kimball, R.; and Caserta, J. “The Data Warehouse ETL Toolkit: Practical Techniques for Extracting and Cleaning”. John Wiley & Sons, Inc. New York, USA. 2004.

[44] da COSTA, A. M. P. M. “A Gestão da Qualidade dos Dados em Ambientes de Data Warehousing na Prossecução da Excelência da Informação,” Dissertação (Mestrado em Sistemas de Dados e Processamento Analítico), Universidade do Minho, Portugal, 2006.

[45] Ciferri, C.; Souza, F. “Focusing on Data Distribution in the WebD2W System,” In *Proceedings of the 4th International Conference on DW and Knowledge Discovery*, pp. 265-274, Aix-en-Provence, France, vol. 2454 of Lecture Notes in Computer Science, Springer, 2002.

[46] McKnight, D. H.; and Chervany, N. L. “The Meanings of Trust”, *tech. report MISRC Working Paper Series 96-04*, Management Information Systems Research Center, University of Minnesota, 1996.

[47] Jøsang, A.; Ismail, R.; and Boyd, C. “A Survey of Trust and Reputation Systems for Online Service Provision,” *Decision Support Systems*, vol. 43, no. 2, pp. 618-644, 2007.

[48] Gambetta, D. “Trust: Making and Breaking Cooperative Relations”, Basil Blackwell, 1988.

[49] Fan, M.; Tan, Y.; and Whinston, A. B. “Evaluation and Design of Online Cooperative Feedback Mechanisms for Reputation Management,” *IEEE Trans. Data and Knowledge Eng.*, vol. 17, no. 2, pp. 244-254, 2005.

[50] Zacharia, G.; and Maes, P. “Collaborative Reputation Mechanisms in Electronic Marketplaces,” *Proc. 32nd Hawaii Int’l Conf. System Sciences*, IEEE CS Press, 1999.

[51] Resnick et al. “Reputation Systems,” *Comm. ACM*, vol. 43, no.12, pp.45-48, 2000.

[52] Ziegler, C.; and Skubacz, M. “Towards Automated Reputation and Brand Monitoring on the Web,” *Proc. IEEE/WIC/ACM Int’l Conf. Web Intelligence*, IEEE CS Press, pp. 1066-1072, 2006.

[53] Kim, W.; Choi, B.-J.; Hong, E.-K.; Kim, S.-K.; and Lee, D. “A Taxonomy of Dirty Data,” *Data Mining and Knowledge Discovery*, vol. 7, pp. 81-99, 2003.

[54] Oliveira, P.; Rodrigues, F.; Henriques, P.; and Galhardas, H. “A Taxonomy of Data Quality Problems”. In *Proceedings of the 2nd International Workshop on Data and Information Quality (in conjunction with CAiSE’05)*, Porto, Portugal, jun. 2005.

[55] Singh, R; and Singh, K. “A descriptive classification of causes of data quality problems in data warehousing,” *International Journal of Computer Science Issues*, vol. 7, no. 3, pp. 41-50, 2010.

[56] Li, L.; Peng, T.; Kennedy, J. “A Rule Based Taxonomy of Dirty Data,” In *Proceedings of Annual International Academic Conference on Data Analysis, Data Quality and Metadata Management*. GSTF, Singapure. vol.1, no. 2, pp. 140-148, 2010.

[57] Heinrich, B.; Kaiser, M.; and Klier, M. “How to measure data quality? - A metric based approach,” In *Proceedings of the 28th International Conference on Information Systems (ICIS)*, 2007.

[58] Heinrich, B.; Klier, M. “Assessing data currency – a probabilistic approach,” *Journal of Information Science,* vol. 37, no. 1, pp. 86-100, 2011.

[59] Even, A.; Shankaranarayanan, G. “Utility-driven assessment of data quality,” *The DATA BASE for Advances in Information Systems*, vol. 38, no. 2, p. 75-93, 2007.

[60] Ballou, D.P.; Wang, R. Y.; Pazer, H. L.; Tayi, G. K. “Modeling information manufacturing systems to deter-mine information product quality,” *Management Science*, vol. 44, no. 4, p. 462-484, 1998.

[61] Heinrich, B.; Kaiser, M.; and Klier, M. “A procedure to develop metrics for currency and its application in CRM,” *ACM Journal of Data and Information Quality*, vol. 1, no. 1, pp. 5-28, 2009.

[62] Geiger, J. “Data Quality Management: The Most Critical Initiative You Can Implement”, Intelligent Solutions, Inc., Boulder, Paper 098-29, 2004.

[63] Kook, Y.-G.; Lee, J.; Park, M.-W; Choi, K.-S.; Kim, J.-S.; Shin S.-S. “Data Quality Management Based on Data Profiling in E-Government Environments,” *Advanced Communication and Networking Communications in Computer and Information Science*, vol. 99, pp. 286-291, 2011.

[64] Boyadzhieva D.; Kolev B. “An Extension of the Relation Model to Intuitionistic Fuzzy Data Quality Attribute Model,” In *Proceedings of 4th International IEEE Conference on Intelligent Systems*, vol. 2, pp. 13-19, 2008.

[65] Redman, T. “The Impact of Poor Data Quality on the Typical Enterprise,” *Communications of the ACM,* vol. 41, no. 2, pp. 79-82, 1998.

[66] Oliveira, P. J. “Detecção e Correcção de Problemas de Qualidade dos Dados: Modelo, Sintaxe e Semântica”, Ph.D. thesis, Universidade do Minho - Escola de Engenharia, Portugal, 2008.

[67] Rahm, E.; and Do, H. H. “Data Cleaning: Problems and Current Approaches,” *Bulletin of the Technical Committe on Data Engineering – Special Issue on Data Cleaning*, vol. 23, no. 4, pp. 3-13, 2000.

[68] Müller, H.; and Freytag, J.-C. “Problems, Methods, and Challenges in Comprehensive Data Cleansing,” *Technical Report HUB-IB-164*, Berlin: Humboldt-Universitt zu Berlin, Institut fr Informatik, 2003.

[69] Hassine-Guetari, S. B.; Darmont, J.; Chauchat, J.-H. “Aggregation of data quality metrics using the Choquet integral,” *Proceedings of the 8th International Workshop on Quality in Databases*, 2010.

[70] Zadeh, L. A. “Fuzzy Sets,” *Information and Control*, vol. 8, no. 3, pp. 338-353, 1965.

[71] Maydanchik, A. “Causes of Data Quality Problems, Data Quality Assessment,” *Techniques Publications LLC*, Capítulo 1, 2007. Disponível em: <http://media.techtarget.com/searchDataManagement/downloads/Data_Quality_Assessment_-_Chapter_1.pdf>.

[72] Etcheverry, L.; Peralta, V.; and Bouzeghoub, M. “Qbox-Foundation: a Metadata Platform for Quality Measurement”. In *Proceeding of the 4th Workshop on Data and Knowledge Quality (DKQ’2008)*, Nice, France, 2008.

[73] Bertino, E.; Dai, C.; and Kantarcioglu, M. “The challenge of assuring data trustworthiness,” In *DASFAA ’09: Proceedings of the 14th International Conference on Database Systems for Advanced Applications*, (Berlin, Heidelberg), pp. 22-33, Springer-Verlag, 2009.

[74] Prat, N.; Madnick, S. “Measuring Data Believability: A Provenance Approach”. In *Proc. of the 41st Hawaii Int. Conference on System Sciences (HICSS)*, 2008.

[75] Ross, J. T. “Fuzzy Logic with Engineering Applications,” John Wiley & Sons, 3a. edição, 2010.

[76] Staab, S.; Bhargava, B.; Lilien, L.; Rosenthal, A.; Winslett, M.; Sloman, M.; Dillon, T. S.; Chang, E.; Hussain, F. K.; Nejdl, W.; Olmedilla, D.; Kashyap, V. “The pudding of trust,” *IEEE Intell. Syst.*, vol. 19, no. 5, pp. 74-88, 2004.

[77] Sottara, D.; Mello, P.; and Proctor, M. “A Configurable Rete-OO Engine for Reasoning with Different Types of Imperfect Information,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, 2010.

[78] Flaminio, T.; Godo, L.; Marchioni, E. Reasoning about uncertainty of fuzzy events: an overview. Preprint, 2011.

[79] Peralta, V. "Data Freshness and Data Accuracy: a State of the Art," *Technical Report TR13-06*, In.Co., Universidad de la República, Uruguai, March 2006.

[80] Jarke, M.; Jeusfeld, M. A.; Quix, C.; and Vassiliadis, P. "Architecture and quality in Data Warehouses: An extended repository approach," *Information Systems*, vol. 24, no. 3, pp. 229-253, 1999.

[81] Angeles, M. P.; and MacKinnon, L. "Assessing Data Quality of Integrated Data by Quality Aggregation of its Ancestors," *Computación y Sistemas*, vol. 13, no. 3, pp. 331-344, 2010.

[82] Zadeh, L. A. "Is there a need for fuzzy logic?," *Information Sciences*, vol. 178, no. 13, pp. 2751-2779, 2008.

[83] Haug, A.; Pedersen, A.; and Arlbjørn, J.S. "A classification model of ERP system data quality," Industrial Management & Data Systems, vol. 109, no. 8, pp. 1053-1068, 2009. doi:10.1108/02635570910991292

[84] Golfarelli, M. "Open source BI platforms: a functional and architectural comparison". In *DaWaK*, pp. 287-297. Springer, 2009

[85] Sattler, K.-U.; and Schallehn, E. "A Data Preparation Framework Based on a Multidatabase Language," In *Proceedings of the International Database Engineering and Applications Symposium*, Grenoble, France, pp. 219-228, 2001.

[86] Haug, A.; Zachariassen, F.; van Liempd, D. "The costs of poor data quality," *Journal of Industrial Engineering and Management (JIEM 2011)*, vol. 4, no. 2, pp. 168-193, 2011.