



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ONTODW: UMA ABORDAGEM PARA A EXTRAÇÃO DE PERSPECTIVAS DE
ANÁLISE A PARTIR DE *DATA WAREHOUSES*

Tiago Outerelo da Silva

Orientadoras

Prof. Dr.^a Fernanda Araujo Baião Amorim

Prof. Dr.^a Kate Cerqueira Revoredo

Rio de Janeiro, RJ – Brasil

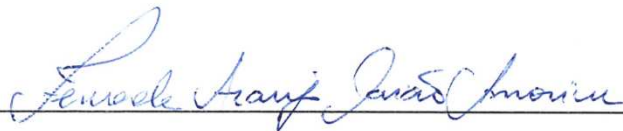
Setembro de 2016

ONTODW: UMA ABORDAGEM PARA A A EXTRAÇÃO DE
PERSPECTIVAS DE ANÁLISE A PARTIR DE *DATA WAREHOUSES*

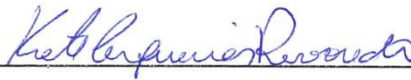
Tiago Outerelo da Silva

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA
OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-
GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO
DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO
EXAMINADORA ABAIXO ASSINADA.

Aprovada por:



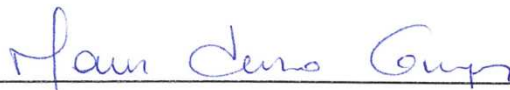
Prof. Fernanda Araujo Baião Amorim, D.Sc – UNIRIO



Prof. Kate Cerqueira Revoredo, D.Sc – UNIRIO



Prof. Astério Kiyoshi Tanaka, Ph. D. – UNIRIO



Prof. Maria Luiza Machado Campos, Ph. D. – UFRJ

Rio de Janeiro, RJ – Brasil

Setembro de 2016

S586 Silva, Tiago Outerelo da.
Ontodw: uma abordagem para a extração de perspectivas de análise a partir de *data warehouses* / Tiago Outerelo da Silva, 2016.
171 f. ; 30 cm

Orientadora: Fernanda Araujo Baião Amorim.
Coorientadora: Kate Cerqueira Revoredo.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

1. Data warehouse. 2. Armazenamento de dados. 3. Inteligência competitiva (Administração). 4. Ontologia. I. Amorim, Fernanda Araujo Baião. II. Revoredo, Kate Cerqueira. III. Universidade Federal do Estado do Rio de Janeiro. Centro de Ciências Exatas e Tecnológicas. Curso de Mestrado em Informática. IV. Título.

CDD – 005.74

“O futuro pertence àqueles que acreditam na beleza de seus sonhos.”

Eleanor Roosevelt

Agradecimentos

Primeiramente, agradeço à minha esposa Lívia pelo apoio e paciência nesse período de estudos, que culminou numa rotina mais corrida e trabalhosa. Rotina essa que encurtou nossos momentos de convivência, especialmente junto ao nosso filho Mateus, de apenas 5 anos. A dificuldade existiu, mas se fez necessária para cumprir esta missão. Agradeço também aos meus pais, irmãos e toda a minha família, pelo estímulo constante e incondicional, sem esquecer é claro, dos amigos, pela importante contribuição nas pesquisas.

Agradeço à UNIRIO pela dedicação e investimento recebidos, essenciais ao longo do curso. Obrigado a todos os professores, principalmente ao professor Tanaka pelos preciosos conselhos fornecidos ao longo do trabalho desenvolvido. Agradeço de forma destacada às minhas orientadoras Fernanda e Kate, pelos ensinamentos, paciência e constante apoio. A dedicação, competência e forma de coordenar os estudos me fizeram aproveitar os novos conhecimentos adquiridos e ter grande satisfação ao avançar no curso. Agradeço ainda aos novos amigos que fiz, em especial ao Valdemar, André, Fernando, Bianca e Rafael, que contribuíram muito para que eu conseguisse concluir este trabalho.

Por fim, agradeço aos colegas de trabalho pelo apoio e incentivo. Em especial, agradeço aos colegas de equipe, que tiveram sobrecarga no trabalho em função da minha ausência na reta final das atividades acadêmicas, e aos meus gestores por acreditar em meu potencial e me concederem um período de estudo em um momento em que necessitava sem nenhum questionamento. Destaco o apoio do meu gestor direto Paulo, por acompanhar de perto minhas dificuldades e sempre se colocar à disposição para me ajudar no que foi necessário, e dos clientes internos que se dispuseram a participar da pesquisa em diversos momentos.

DA SILVA, Tiago Outerelo. **ONTODW: UMA ABORDAGEM PARA A EXTRAÇÃO DE PERSPECTIVAS DE ANÁLISE A PARTIR DE DATA WAREHOUSES**, UNIRIO, 2016. 171 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Business Intelligence (BI) promove a adequada tomada de decisão nas organizações, principalmente fornecendo os meios para analisar dados históricos armazenados em repositórios chamados *Data Warehouses* (DW). No entanto, uma representação formal dos conceitos implementados em um DW raramente existe, o que seria importante para esclarecer e semanticamente descrever os conceitos por trás dos dados armazenados em um DW, bem como os conceitos analíticos que estão disponíveis para as ferramentas de BI. Exemplos de importantes peças de conhecimento que estão frequentemente ocultas no DW são: quais os conceitos que estão disponíveis como perspectivas de análise (dimensões), como os conceitos se relacionam entre si, quais medidas (fatos) estão disponíveis e o que elas significam, que perspectivas de domínio são consideradas para cada medida e como as medidas podem ser agregadas. Por outro lado, um dos usos relevantes de uma ontologia para a área de Ciência da Computação é como um artefato codificado que representa formalmente uma conceitualização compartilhada sobre um universo de discurso. Portanto, ontologias podem ser utilizadas para representar conceitos de domínio e conceitos analíticos codificados e armazenados em um DW. No entanto, extrair esses conceitos a partir de um DW em produção não é uma tarefa trivial, especialmente em organizações de médio e grande porte, frequentemente com dezenas de medidas e dezenas (mesmo centenas) de dimensões e potenciais agregações. Neste trabalho, é definido um conjunto de regras de mapeamento entre construtos de DW e elementos conceituais (conceitos e relacionamentos), no sentido de extrair automaticamente elementos estruturais da aplicação de BI em uma ontologia codificada em OWL. A proposta foi avaliada com sucesso em um cenário real de um fundo de pensão dos funcionários de uma instituição financeira brasileira.

Palavras-chave: *Business Intelligence*, *Data Warehouse*, ontologia.

ABSTRACT

Business Intelligence (BI) fosters proper decision-making in organizations, mainly by providing the means to analyze historical data stored in repositories called Data Warehouses (DW). However, a formal representation of the concepts implemented in a DW rarely exists, which would be important to clarify and semantically describe the concepts behind the data stored in a DW, as well as the analytical concepts that are available for the BI tools. Examples of important pieces of knowledge that are frequently hidden into the DW are: which concepts are available as analysis perspectives (dimensions), how the concepts relate to each other, which metrics (facts) are available and what do they mean, which domain perspectives are considered for each metric and how metrics may be aggregated. On the other hand, one of the relevant uses of an ontology for the Computer Science area is as a codified artifact that formally represents a shared conceptualization about a universe of discourse. Therefore, ontologies can be used to represent both domain and analytical concepts codified and stored in a DW. However, extracting these concepts from an already-in-production DW is not a trivial task, especially in medium and large organizations, often with tens of metrics and tens (even hundreds) of dimensions and potential aggregations. In this dissertation, we define a set of mapping rules from DW constructs to conceptual elements (concepts and relationships), towards automatically extracting structural elements of the BI application in an ontology codified in OWL. The proposal was successfully evaluated in a real scenario of the pension fund of employees of a Brazilian financial institution.

Keywords: Business Intelligence, Data Warehouse, ontology.

Sumário

| | |
|--|----|
| 1 - Introdução..... | 15 |
| 1.1 Motivação | 15 |
| 1.2 Problema | 15 |
| 1.3 Justificativa | 16 |
| 1.4 Proposta de solução | 16 |
| 2 – Fundamentação Teórica..... | 18 |
| 2.1 <i>Business Intelligence</i> | 18 |
| 2.1.1 <i>Data Warehouse</i> | 20 |
| 2.1.2 Modelagem Multidimensional | 23 |
| 2.1.3 OLAP | 28 |
| 2.2 Ontologia | 30 |
| 2.2.1 OWL..... | 33 |
| 2.3 Considerações finais | 34 |
| 3 – Proposta de Solução | 35 |
| 3.1 Regras de Mapeamento..... | 41 |
| 3.1.1 Classe Fact..... | 42 |
| 3.1.2 Classe Dimension | 43 |
| 3.1.3 Classe DimensionLevel..... | 44 |
| 3.1.4 Classe Attribute | 45 |
| 3.1.5 Classe Measure | 46 |
| 3.1.6 Classe Rollup | 47 |
| 3.1.7 Classe Hierarchy | 49 |
| 3.1.8 Classe SummarizabilityAlongFact | 50 |
| 3.1.9 Classe SummarizabilityAlongDimension..... | 51 |
| 3.1.10 Classe SummarizabilityAlongHierarchy..... | 52 |
| 3.2 Transformação para OWL | 53 |
| 3.3 Implementação da solução..... | 54 |
| 4 – Estudo de Caso | 61 |
| 4.1 Cenário de aplicação | 61 |
| 4.2 Projeto de avaliação | 65 |

| | |
|---|-----|
| 4.3 Execução do estudo de caso..... | 68 |
| 5 – Análise de Resultados..... | 71 |
| 5.1 Visão geral da análise dos resultados | 71 |
| 5.2 Análise da ontologia do estudo de caso | 71 |
| 5.3 Análise das pesquisas com especialistas..... | 76 |
| 5.3.1 Perfil dos respondentes..... | 77 |
| 5.3.2 Resultados das regras R1 e R2 | 82 |
| 5.3.3 Resultados das regras R3, R4, R8 e R10..... | 87 |
| 5.3.4 Resultados das regras R7 e R9 | 97 |
| 5.3.5 Resultados das regras R5, R6 e R11 | 100 |
| 5.3.6 Resultados das regras R12 e R13 | 106 |
| 5.3.7 Considerações sobre os resultados | 111 |
| 5.4 Análise da utilidade da ontologia gerada por usuários | 112 |
| 5.4.1 Reunião prévia com os usuários..... | 113 |
| 5.4.2 Perfil..... | 114 |
| 5.4.3 Resultados | 115 |
| 5.4.4 Considerações sobre os resultados | 122 |
| 6 – Trabalhos relacionados | 124 |
| 7 – Conclusão | 128 |
| Referências | 132 |
| ANEXO I – Questionário 1 para profissionais de BI do mercado | 134 |
| ANEXO II – Questionário 2 para profissionais de BI da organização do estudo de caso | 144 |
| ANEXO III – Questionário 3 para usuários do sistema de BI | 157 |
| ANEXO IV – Respostas subjetivas do Questionário 2 | 164 |
| ANEXO V – Respostas subjetivas do Questionário 3..... | 168 |

Lista de Figuras

| | |
|---|----|
| Figura 2.1: Exemplo de arquitetura de <i>Business Intelligence</i> [Chaudhuri, Dayal and Narasayya, 2011] | 19 |
| Figura 2.2: Arquitetura CIF proposta por Bill Inmon [Ross, 2004] | 21 |
| Figura 2.3: Arquitetura de barramento proposta por Ralph Kimball [Ross, 2004] | 21 |
| Figura 2.4: Exemplo de modelo de dados do tipo estrela..... | 24 |
| Figura 2.5: Exemplo de modelo de dados do tipo floco de neves | 26 |
| Figura 2.6: Exemplo de modelo de dados com tabela de fato sem fato | 27 |
| Figura 2.7: Exemplo de visão multidimensional (cubo) de dados | 28 |
| Figura 2.8: Exemplo de operação de <i>roll up</i> | 29 |
| Figura 2.9: Exemplo de operação de <i>dice</i> | 29 |
| Figura 2.10: Classificação de ontologias | 31 |
| Figura 2.11: Exemplo de ontologia de domínio | 32 |
| Figura 2.12: Relação entre sub-linguagens OWL | 34 |
| Figura 3.1: Visão geral da geração de ontologia | 36 |
| Figura 3.2: Metamodelo de tarefa OLAP [Prat, Megdiche and Akoka, 2012]..... | 37 |
| Figura 3.3: Diagrama do esquema de dados multidimensional do domínio de cadastro de funcionários | 41 |
| Figura 3.4: Regra R1 - Mapeamento de conceitos Fact | 42 |
| Figura 3.5: Regra R2 - Mapeamento de conceitos Dimension | 44 |
| Figura 3.6: Regra R3 - Mapeamento de conceitos DimensionLevel | 44 |
| Figura 3.7: Regra R4 - Mapeamento de conceitos Attribute | 46 |
| Figura 3.8: Regra R5 - Mapeamento de conceitos Measure | 47 |
| Figura 3.9: Regra R6 - Mapeamento de conceitos Measure | 47 |
| Figura 3.10: Regra R7 - Mapeamento de conceitos RollUp | 48 |
| Figura 3.11: Regra R8 - Mapeamento de conceitos RollUp | 48 |
| Figura 3.12: Regra R9 - Mapeamento de conceitos Hierarchy | 50 |
| Figura 3.13: Regra R10 - Mapeamento de conceitos Hierarchy | 50 |
| Figura 3.14: Regra R11 - Mapeamento de conceitos SummarizabilityAlongFact | 51 |
| Figura 3.15: Regra R12 - Mapeamento de conceitos SummarizabilityAlongDimension | 51 |

| | |
|---|----|
| Figura 3.16: Regra R13 - Mapeamento de conceitos SummarizabilityAlongHierarchy | 52 |
| Figura 3.17: Exemplo de transformação para classe na ontologia OWL [Prat, Megdiche and Akoka, 2012] | 53 |
| Figura 3.18: Diagrama de componentes da implementação do OntoDW | 55 |
| Figura 3.19: Diagrama de sequência da execução do OntoDW | 56 |
| Figura 3.20: Arquitetura do ambiente de desenvolvimento | 57 |
| Figura 3.21: Recorte da ontologia extraída do ambiente de testes | 60 |
| Figura 4.1: Recorte do modelo do esquema do DW do estudo de caso com dimensões relacionadas ao conceito Funcionário | 62 |
| Figura 4.2: Recorte do modelo do esquema do DW do estudo de caso com fatos e dimensões | 63 |
| Figura 4.3: Representação das tabelas FAT_FUNCIONARIO e AGR_FUNCIONARIO_3 | 64 |
| Figura 5.1: Captura de tela do Protégé com instâncias da ontologia..... | 72 |
| Figura 5.2: Recorte da ontologia extraída no estudo de caso | 73 |
| Figura 5.3: Níveis de dimensão da tabela DIM_FUNCIONARIO e atributos correspondentes | 74 |
| Figura 5.4: Perfil do grupo 1 em relação a modelagem multidimensional..... | 77 |
| Figura 5.5: Perfil do grupo 1 em relação a aplicações OLAP | 78 |
| Figura 5.6: Perfil do grupo 1 em relação à sua experiência | 78 |
| Figura 5.7: Perfil do grupo 2 em relação a modelagem multidimensional..... | 79 |
| Figura 5.8: Perfil do grupo 2 em relação a aplicações OLAP | 80 |
| Figura 5.9: Perfil do grupo 2 em relação à sua experiência | 80 |
| Figura 5.10: Perfil do grupo 2 em relação a tempo de empresa | 81 |
| Figura 5.11: Recorte 1 do DW utilizado para validar as regras R1 e R2 | 82 |
| Figura 5.12: Questão para validar as regras R1 e R2 | 83 |
| Figura 5.13: Questões subjetivas sobre a seção 4 do formulário para o grupo 2 | 83 |
| Figura 5.14: Respostas do grupo 1 para as tabelas de dimensão identificadas pelo OntoDW através da regra R1..... | 84 |
| Figura 5.15: Respostas do grupo 2 para as tabelas de dimensão identificadas pelo OntoDW através da regra R1..... | 84 |
| Figura 5.16: Respostas do grupo 1 para as tabelas de fato identificadas pelo OntoDW através da regra R2 | 85 |
| Figura 5.17: Respostas do grupo 2 para as tabelas de fato identificadas pelo OntoDW através da regra R2 | 86 |
| Figura 5.18: Recorte 2 do DW utilizado para validar as regras R3, R4, R8 e R10..... | 88 |
| Figura 5.19: Questões para validar as regras R3, R8, R10 e R4 | 88 |

| | |
|---|-----|
| Figura 5.20: Questões subjetivas sobre a seção 5 do formulário para o grupo 2 | 88 |
| Figura 5.21: Respostas do grupo 1 para a questão 02 | 89 |
| Figura 5.22: Respostas do grupo 2 para a questão 02 | 89 |
| Figura 5.23: Respostas do grupo 1 para a questão 03 | 91 |
| Figura 5.24: Respostas do grupo 2 para a questão 03 | 91 |
| Figura 5.25: Respostas do grupo 1 para a questão 04 | 93 |
| Figura 5.26: Respostas do grupo 2 para a questão 04 | 93 |
| Figura 5.27: Respostas do grupo 1 para a questão 05 | 94 |
| Figura 5.28: Respostas do grupo 2 para a questão 05 | 94 |
| Figura 5.29: Recorte 3 do DW utilizado para validar as regras R7 e R9 | 97 |
| Figura 5.30: Questões para validar as regras R7 e R9..... | 97 |
| Figura 5.31: Questões subjetivas sobre a seção 6 do formulário para o grupo 2 | 97 |
| Figura 5.32: Respostas do grupo 1 para a questão 06 | 98 |
| Figura 5.33: Respostas do grupo 2 para a questão 06 | 98 |
| Figura 5.34: Respostas do grupo 1 para a questão 07 | 99 |
| Figura 5.35: Respostas do grupo 2 para a questão 07 | 100 |
| Figura 5.36: Recorte 4 do DW utilizado para validar as regras R5, R6 e R11..... | 101 |
| Figura 5.37: Questões para validar as regras R5, R6 e R11 | 101 |
| Figura 5.38: Questões subjetivas sobre a seção 7 do formulário para o grupo 2 | 101 |
| Figura 5.39: Respostas do grupo 1 para as questões 08 e 09..... | 102 |
| Figura 5.40: Respostas do grupo 2 para as questões 08 e 09..... | 102 |
| Figura 5.41: Respostas do grupo 1 para a questão 10 | 104 |
| Figura 5.42: Respostas do grupo 2 para a questão 10 | 104 |
| Figura 5.43: Respostas do grupo 1 para a questão 11 | 105 |
| Figura 5.44: Respostas do grupo 2 para a questão 11 | 105 |
| Figura 5.45: Recorte 5 do DW utilizado para validar as regras R12 e R13 | 106 |
| Figura 5.46: Questões para validar as regras R12 e R13..... | 107 |
| Figura 5.47: Questões subjetivas sobre a seção 8 do formulário para o grupo 2 | 107 |
| Figura 5.48: Respostas do grupo 1 para a questão 12 | 108 |
| Figura 5.49: Respostas do grupo 2 para a questão 12 | 108 |
| Figura 5.50: Respostas do grupo 1 para a questão 13 | 109 |
| Figura 5.51: Respostas do grupo 2 para a questão 13 | 110 |
| Figura 5.52: Perfil do grupo 3 em relação ao domínio em conceitos de aplicações de BI | 114 |
| Figura 5.53: Perfil do grupo 3 em relação à sua experiência | 115 |

| | |
|---|-----|
| Figura 5.54: Questões sobre a utilidade de representações de conhecimento | 116 |
| Figura 5.55: Recorte 1 da ontologia para a pesquisa com usuários..... | 117 |
| Figura 5.56: Questões sobre o recorte 1 da ontologia | 117 |
| Figura 5.57: Recorte 2 da ontologia para a pesquisa com usuários..... | 118 |
| Figura 5.58: Questões sobre o recorte 2 da ontologia | 119 |
| Figura 5.59: Recorte 3 da ontologia para a pesquisa com usuários..... | 120 |
| Figura 5.60: Questões sobre o recorte 3 da ontologia | 120 |
| Figura 5.61: Recorte 4 da ontologia para a pesquisa com usuários..... | 121 |
| Figura 5.62: Questões sobre o recorte 4 da ontologia | 122 |

Lista de Tabelas

| | |
|---|-----|
| Tabela 3.1: Exemplo de registro da tabela DIM_TEMPO..... | 45 |
| Tabela 3.2: Níveis de dimensão mapeados do DW de teste..... | 57 |
| Tabela 3.3: Agregabilidades de medidas com fatos mapeadas do DW de teste..... | 58 |
| Tabela 3.4: Agregabilidades de medidas com dimensões mapeadas do DW de teste ... | 58 |
| Tabela 3.5: Agregabilidades de medidas com hierarquias mapeadas do DW de teste... | 59 |
| Tabela 4.1: Quantidade de instâncias de classe mapeadas no estudo de caso..... | 69 |
| Tabela 4.2: Quantidade de relações entre instâncias mapeadas no estudo de caso..... | 69 |
| Tabela 5.1: Níveis de dimensão da tabela FUNCIONARIO | 90 |
| Tabela 5.2: Operações de <i>roll up</i> da dimensão FUNCIONARIO..... | 92 |
| Tabela 5.3: Níveis da dimensão FUNCIONARIO e atributos correspondentes..... | 95 |
| Tabela 6.1: Trabalhos relacionados | 126 |

1 - Introdução

Este capítulo apresenta os principais aspectos desta pesquisa, incluindo a sua motivação, a caracterização problema, a justificativa, a hipótese, a proposta de solução e a estrutura do documento.

1.1 Motivação

Organizações estão sobrecarregadas com a quantidade crescente de dados, que são continuamente gerados e armazenados em repositórios corporativos, a serem analisados para uma adequada tomada de decisão [Sidorova and Torres, 2014]. Definições de estratégias de negócios, decisões sobre os preços dos produtos e tendências de comportamento do cliente são exemplos de cenários que se beneficiam desta análise de dados [Andoh-Baidoo et al., 2014]. Soluções de Inteligência de Negócio (*Business Intelligence*, ou BI) fornecem os meios para recolher informações e para derivar conhecimento através de ferramentas de análise para a tomada de decisão [Sell et al., 2011]. Tais soluções ajudam na análise de grandes volumes de dados, transformando-os em informação significativa, útil e esclarecedora.

1.2 Problema

Apesar da importância de ferramentas analíticas fornecidas por soluções de BI para as organizações atuais, existem desafios para alavancar o seu impacto no processo de tomada de decisão [Sell et al., 2011]. Os usuários (analistas de negócio e analistas de BI) não têm uma definição clara de todas as informações à sua disposição, nem mesmo das possíveis relações entre os dados disponíveis. Isso pode ocorrer devido à falta de integração entre os artefatos de semântica de negócio e as ferramentas analíticas [Sell et al., 2011].

1.3 Justificativa

A percepção é de que os analistas de negócios não utilizam todo o potencial das ferramentas para a realização de análises sobre o negócio e os analistas de BI nem sempre possuem documentação atualizada sobre os conceitos implementados no DW, dificultando o apoio aos analistas de negócios e a realização de manutenções corretivas e evolutivas. No sentido desta integração, uma representação formal pode ser usada para descrever semanticamente os conceitos implementados na solução de BI.

Por outro lado, uma utilização relevante de ontologias para a área de Ciência da Computação é como um artefato que formalmente representa uma conceituação compartilhada de um domínio do discurso, tarefa ou aplicação, através de seus elementos fundamentais: conceitos e relacionamentos. Portanto, é um artefato natural para descrever a semântica por trás dos dados e metadados armazenados em um DW, proporcionando assim uma rica, explícita e atualizável representação dos dados organizacionais.

No entanto, a maioria dos sistemas de informação atuais não apresenta uma ontologia que descreva suas informações de negócio disponíveis, pelo fato de que sua construção não faz parte do ciclo de vida de desenvolvimento de software e pelo processo manual de construção de ontologias ser uma tarefa difícil, dispendiosa, demorada e requerer profundo conhecimento do domínio [El Idrissi, Baïna and Baïna, 2013]. Em particular, sistemas de *Business Intelligence* também são sistemas de informação e apresentam a mesma questão de ausência de uma representação formal de conhecimento que explicita e descreva semanticamente os dados e metadados armazenados no *Data Warehouse*.

1.4 Proposta de solução

Para solucionar este problema, neste trabalho foi definido um conjunto de regras de mapeamento de construtos estruturais de DW para elementos conceituais (conceitos e relacionamentos), com o objetivo de extração automática de uma ontologia codificada

em OWL, no âmbito dos sistemas de BI. A proposta foi avaliada com sucesso através de um estudo de caso em um cenário real.

Para a definição desta proposta de solução, foi utilizada a hipótese que, se definirmos regras de mapeamento específicas a partir dos dados e metadados de Data Warehouses, então é possível extrair ontologias com as perspectivas de análise implementadas no banco de dados (que sejam aderentes aos conceitos de negócio e de BI). Assim, caso a abordagem não possibilite a geração de uma ontologia a partir de um DW, ou que gere uma ontologia sem as perspectivas de análise existentes no esquema de dados, a hipótese não será verdadeira.

Esta dissertação é organizada da seguinte forma:

- Capítulo 2: apresenta a base teórica do trabalho;
- Capítulo 3: apresenta a proposta de solução para o problema caracterizado;
- Capítulo 4: descreve o estudo de caso realizado, incluindo o cenário de aplicação, o projeto de avaliação e detalhes da implementação do OntoDW e de sua execução;
- Capítulo 5: apresenta os resultados obtidos no estudo de caso e a avaliação das pesquisas aplicadas;
- Capítulo 6: descreve os trabalhos encontrados relacionados a geração automática de ontologias a partir de bancos de dados relacionais;
- Capítulo 7: contém as considerações finais.

2 – Fundamentação Teórica

Este capítulo apresenta a fundamentação teórica. Em primeiro lugar, os fundamentos de Business Intelligence, Data Warehouse, modelagem multidimensional e OLAP são introduzidos. Em seguida, os fundamentos de ontologia e OWL.

2.1 Business Intelligence

Em 1958, o termo *Business Intelligence* foi empregado por Luhn [Luhn, 1958] para qualificar um sistema para automatizar o processamento de documentos e distribuir na organização as informações para as pessoas devidas. Desde então, BI tem sido citado de forma frequente no âmbito de sistemas de informação, evoluindo em sua arquitetura e ganhando cada vez mais importância nas organizações. Aplicações de BI são utilizadas em organizações de diversas áreas de atuação, de forma a apoiar a tomada de decisão. Como exemplos, podemos encontrar seu uso em bancos, para identificação de fraudes, em empresas de varejo, para identificação de perfis dos clientes, e em indústrias, para aperfeiçoamento do processo produtivo.

De forma geral, qualquer sistema de informação que auxilie na identificação e resolução de problemas e na tomada de decisões em organizações pode ser definido como Sistema de Apoio à Decisão (DSS) [Power, 2008]. Um DSS orientado a dados enfatiza a manipulação e o acesso de séries históricas dos dados de uma organização, além de dados externos que lhes sejam úteis. O mais alto nível deste tipo de DSS é a implementação de um sistema de *Business Intelligence*, na sua forma mais tradicional: a disponibilização de um *Data Warehouse* (DW), juntamente com um sistema OLAP (*On-Line Analytical Processing*) o acessando [Negash, 2004].

Com a diminuição do custo das tecnologias, a necessidade de diminuição do tempo entre a aquisição das informações e a tomada de decisões, o aumento da competitividade entre as organizações e o aumento da quantidade e diversidade dos dados a serem tratados, as aplicações de BI se tornaram mais que apenas um tipo de DSS, passando a englobar outros tipos de tecnologias [Airinei and Homocianu, 2009].

Assim sendo, podemos definir *Business Intelligence* como um conjunto de teorias, metodologias, arquiteturas e tecnologias com o objetivo de recuperar e transformar dados brutos em informações significativas e úteis, permitindo que os níveis operacional, tático e estratégico de uma organização possam tomar decisões melhores e de forma mais ágil [Airinei and Homocianu, 2009].

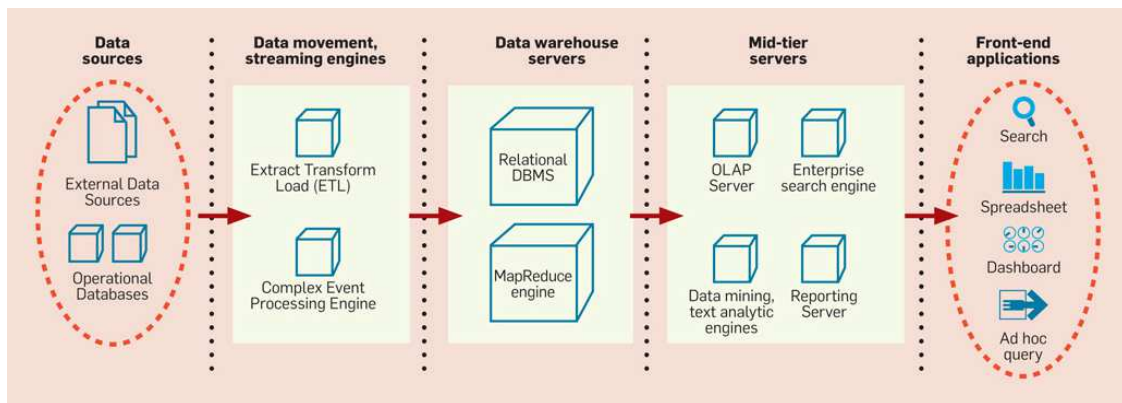


Figura 2.1: Exemplo de arquitetura de *Business Intelligence* [Chaudhuri, Dayal and Narasayya, 2011]

Na Figura 2.1 é mostrada uma arquitetura típica de *Business Intelligence* numa organização, composta de diversas etapas. Os dados são obtidos a partir de fontes diversas, tratados para limpeza e unificação de conceitos, carregados em estruturas específicas para dados analíticos, processados para disponibilização e apresentados aos usuários em interfaces diversas.

As diferentes fontes contêm dados de qualidade, quantidade e forma variados, vindo de fontes externas ou internas e disponíveis em arquivos, serviços ou bancos de dados relacionais, fazendo da etapa de carga dos dados a mais desafiadora de uma implementação de um sistema de BI [Chaudhuri, Dayal and Narasayya, 2011].

Na etapa de carga de dados, temos basicamente dois tipos de processos: o processo ETL (*Extract-Transform-Load*) e o processo CEP (*Complex Event Processing*). O ETL é largamente utilizado e realiza o processo de extração e transformação dos dados para carga em *Data Warehouse*. ETL se refere a um conjunto de ferramentas que desempenham um papel crucial para ajudar a descobrir e corrigir problemas de qualidade de dados e carregar eficientemente grandes volumes de dados para o DW.

A pressão competitiva das empresas de hoje tem levado ao aumento da necessidade de reduzir a latência entre o momento em que os dados operacionais são adquiri-

dos e quando a análise sobre os dados é realizada. O CEP é uma classe de sistemas que aborda esta questão, implementando o que pode ser chamado de BI operacional ou BI em tempo real. BI operacional é diferente do BI tradicional, uma vez que os dados operacionais não necessitam ser carregados pela primeira vez em um DW antes que possam ser analisados [Chaudhuri, Dayal e Narasayya, 2011]. Ele é utilizado em caso de necessidade de consumo das informações em tempo real, e as informações são disponibilizadas imediatamente para consumo, ao invés de somente serem carregadas em um *Data Warehouse*.

Servidores de *Data Warehouse* podem ser divididos em dois tipos: bancos de dados relacionais e bancos de dados não relacionais (NoSQL). Os bancos de dados relacionais são mais comumente utilizados para implementação de *Data Warehouse*. Os bancos de dados não relacionais são utilizados em situações onde é necessário armazenar informações não estruturadas, como informações obtidas em redes sociais, por exemplo.

2.1.1 *Data Warehouse*

De acordo com Inmon [Inmon, 2002], podemos definir um *Data Warehouse* como uma coleção de dados, orientada a assunto, não volátil, integrada e variante no tempo para apoio na tomada de decisão. Um DW é orientado a assunto, pois os dados relacionam eventos ou objetos da vida real, não volátil, pois os dados não são atualizados ou deletados, integrados, pois consiste informações de diversas fontes distintas, e variante no tempo, pois os dados são apresentados com visões históricas.

Um *Data Warehouse* é construído integrando as informações dos processos de negócio da organização, a partir de diversas fontes de informação e com a realização de cargas periódicas. Ele é estruturado para priorização da leitura, frente à inserção ou atualização, e para armazenamento de grandes quantidades de dados, inseridos normalmente através de cargas em lote. Quando se aborda o tema *Data Warehouse*, William H. Inmon e Ralph Kimball são automaticamente lembrados, devido à importância de contribuições que fizeram na área.

William H. Inmon, ou Bill Inmon, é conhecido como o pai do *Data Warehouse*, sendo quem definiu o termo na década de 1970. Na arquitetura CIF (*Corporate Information Factory*), também conhecida como arquitetura Inmon, o *Data Warehouse* é um repositório de informações corporativas com modelo de dados normalizado e é constru-

Ído integrando as informações dos processos de negócio da organização num modelo relacional, a partir de diversas fontes de informação e com cargas periódicas.

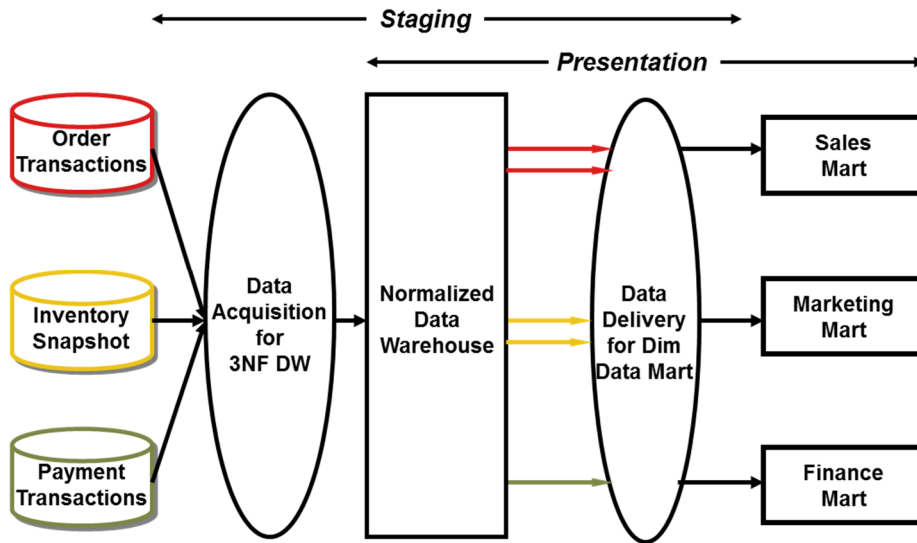


Figura 2.2: Arquitetura CIF proposta por Bill Inmon [Ross, 2004]

Bill Inmon também define que os dados dos processos de negócios são integrados e armazenados no DW mantendo o nível de detalhe original das fontes de origem. O detalhamento de uma informação é também chamado de granularidade, ou simplesmente grão. Ao se consolidar uma informação por alguma perspectiva, aumentamos o grão dessa informação. Ex.: Ao obtermos a informação de valor de vendas de uma determinada rede de lojas, cada registro representa uma venda de um produto realizada por um vendedor. A consolidação das vendas por filial representaria um aumento do grão da informação de valor das vendas, de vendedor para filial.

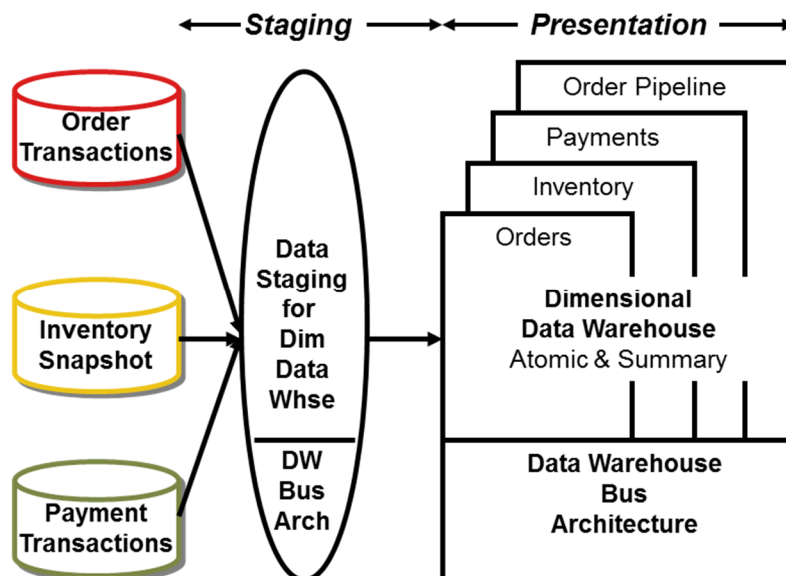


Figura 2.3: Arquitetura de barramento proposta por Ralph Kimball [Ross, 2004]

Na arquitetura CIF, o DW é também utilizado para a realização de cargas em *Data Marts*. Um *Data Mart* é uma coleção de dados constituída a partir de um subconjunto dos dados do *Data Warehouse*. Os dados são carregados para um *Data Mart* de acordo com assuntos específicos para um determinado grupo ou departamento de uma organização, normalmente com maior granularidade. Os *Data Marts* apresentam um conjunto de informações e granularidade próprios e representam visões dos dados armazenados no *Data Warehouse*. Na Figura 2.2 é mostrada uma arquitetura típica de *Data Warehouse* proposta por Bill Inmon.

Ralph Kimball é um pesquisador da área de *Data Warehouse* e *Business Intelligence*. Ele defende que o *Data Warehouse* deve ser projetado para ser rápido e facilmente compreendido [Kimball, 1997]. Na arquitetura de barramento, também conhecida como arquitetura Kimball, o *Data Warehouse* pode ser descrito como o conjunto de todos os *Data Marts*. Os *Data Marts* são implementados com modelo de dados desnormalizado, no menor grão necessário, e disponibilizam informações aos usuários finais para análises e tomada de decisão. As informações são carregadas a partir das fontes de dados diretamente para os *Data Marts*, respeitando a necessidade de atualização e granularidade de cada um. Na Figura 2.3 é mostrada uma arquitetura típica de *Data Warehouse* proposta por Ralph Kimball.

Independente da abordagem adotada (arquitetura CIF, arquitetura barramento ou uma arquitetura híbrida) um *Data Warehouse* armazena informações com o objetivo de prover análise sobre assuntos e apoiar tomadas de decisão, obtidas através da integração de fontes de dados diversas e heterogêneas. Entretanto, existem importantes diferenças entre as arquiteturas, das quais podemos citar as seguintes como principais: necessidade de *Data Warehouse* integrado, abordagem de implementação do *Data Warehouse* na organização e modelo de dados utilizado no *Data Warehouse*.

Segundo a arquitetura CIF, existe a necessidade de implementação de um *Data Warehouse* integrando os dados dos processos de negócio da organização, construindo assim uma “versão única da verdade”. A partir deste *Data Warehouse* integrado, são realizadas cargas para *Data Marts* de assuntos específicos e com dados consolidados, para uso pelos usuários finais, mas mantendo os dados do DW íntegros. A arquitetura de barramento defende que os dados lidos das fontes de dados sejam disponibilizados diretamente nos *Data Marts*, que mantêm apenas uma cópia das informações transacionais, mas modelado por assunto. Não existe a necessidade de um *Data Warehouse* inte-

grado e as informações podem ser armazenadas nos *Data Marts* com replicação e sem granularidades divergentes.

As arquiteturas também divergem quanto à abordagem de implementação do *Data Warehouse*. A abordagem de Inmon é conhecida como top-down, onde primeiro deve ser construído o *Data Warehouse* integrado, para depois serem disponibilizados *Data Marts* para utilização pelos usuários. A abordagem de Kimball é conhecida como bottom-up, onde os *Data Marts* são carregados e disponibilizados aos usuários. O *Data Warehouse* vai sendo ampliado na proporção em que mais *Data Marts* vão sendo criados.

Bill Inmon apresenta uma arquitetura mais robusta, buscando atender a toda a organização desde a implementação inicial. O *Data Warehouse* integrado, além de servir de fonte de dados para os *Data Marts*, pode ter outras funções, como consulta direta pelos usuários e checagem qualidade dos dados de sistemas transacionais. No entanto, o projeto do *Data Warehouse* tende a ser longo, devido à duplicação de ambientes que é proposta e à abordagem top-down, e a carga dos dados sumarizados dos *Data Marts* pode causar dependência excessiva dos usuários à área de TI, pois se uma determinada visão necessária não foi prevista, a carga dos *Data Marts* tem que ser revista.

Ralph Kimball apresenta uma arquitetura mais simplificada, focando na agilidade da disponibilização dos *Data Marts* e rapidez na consulta das informações no banco de dados. Os projetos tendem a ter entregas mais rápidas, disponibilizando *Data Marts* à organização sequencialmente, crescendo o *Data Warehouse* de forma incremental. No entanto, com as informações dispostas de forma descentralizada, é maior o risco de existir uma mesma informação no DW com regras de negócio distintas ou com mesma regra de negócio, mas valores diferentes por conta de problema relativo à carga de dados.

Por último, as abordagens divergem no modelo de dados utilizados para criação do *Data Warehouse*. Enquanto Inmon defende que deve ser utilizado um modelo de dados normalizado, Kimball defende a utilização de um modelo de dados desnormalizado.

2.1.2 Modelagem Multidimensional

A modelagem multidimensional é uma técnica de modelagem de dados orientada a assuntos, amplamente utilizada em ambientes de BI no projeto de *Data Warehouse*.

ses (DW). Os elementos básicos dos modelos multidimensionais são os fatos e dimensões. Tipicamente, os fatos apresentam as informações que se quer medir e são valores numéricos, que podem ser agregados, e as dimensões apresentam as visões a partir das quais se quer analisar as informações e são valores descritivos [Kimball, 1997].

Fatos e dimensões são armazenados em diferentes tabelas em um esquema de dados relacional. Uma tabela de fato pode conter diversos fatos, representados em uma ou mais colunas da tabela. Da mesma forma, uma tabela de dimensão pode conter mais de uma dimensão, representadas em uma ou em um grupo de colunas. Uma tabela de fato também pode ser chamada de tabela de agregação, caso possua os mesmos fatos que outra tabela e apresente uma granularidade maior que a dela. Para a realização do aumento de granularidade de um fato, é preciso que seja definida uma função a ser utilizada para o cálculo, chamada de função de agregação. Como exemplos de função de agregação, podemos citar a função de soma, de média ou de contagem.

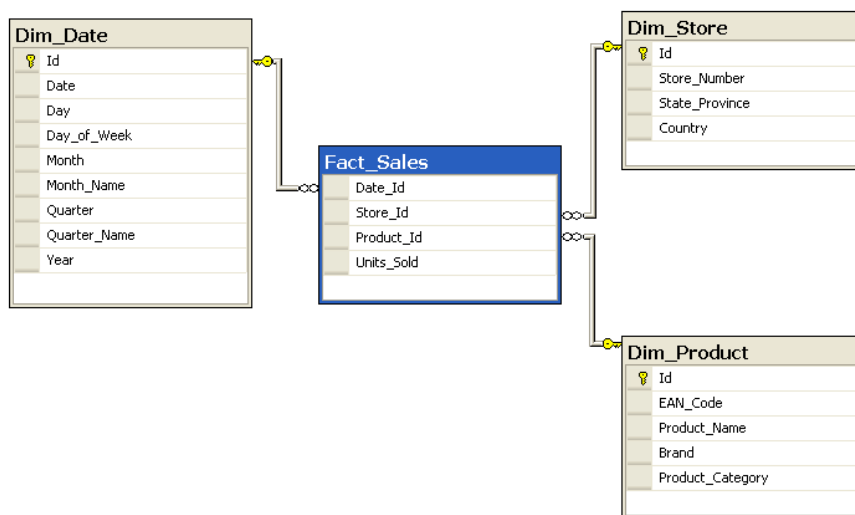


Figura 2.4: Exemplo de modelo de dados do tipo estrela

Sob a perspectiva de aumento de granularidade, um fato pode ser de 3 tipos: aditivo, semi-aditivo ou não aditivo. Um fato aditivo pode ser agregável por qualquer uma das dimensões relacionadas a ele, um fato semi-aditivo pode ser agregável por uma parte das dimensões relacionadas a ele e um fato não aditivo não pode ser agregável por nenhuma das dimensões relacionadas a ele. Sob a perspectiva da natureza do dado armazenado, um fato pode também ser de 3 tipos: estoque, fluxo ou valor por unidade. Um fato do tipo estoque armazena dados momentâneos, representando “fotografias” de um dado em determinados períodos, como o estoque de produtos de uma loja por exemplo. Um fato do tipo fluxo armazena ocorrências de valores de um dado ao longo do

tempo, como as vendas de produtos de uma loja por exemplo. Um fato do tipo valor por unidade armazena dados que comumente são utilizados somente na granularidade em que estão armazenados, como o ranking de produtos mais vendidos de uma loja por exemplo.

Kimball [Kimball, 1997] defende que o *Data Warehouse* deve ser projetado para ser rápido e facilmente compreendido, implementado com modelo de dados desnormalizado e no menor grão necessário. O tipo de modelagem multidimensional criado por Ralph Kimball foi chamado de *Star Schema*, ou modelo estrela. No modelo estrela, as tabelas de dimensão podem apresentar mais de uma dimensão / visão de análise, que são representadas por grupos de colunas. Além disso, a tabela de fato sempre apresenta o nível mais detalhado da informação analisada. Na Figura 2.4 é ilustrado um exemplo de modelo do tipo estrela que armazena dados para a análise de vendas de produtos de uma rede de lojas. O fato armazenado na tabela de fato *Fact_Sales* é o de unidades vendidas de determinado produto, representado pela coluna *Units_Sold*. As tabelas restantes do modelo são tabelas de dimensão e cada uma delas representa um detalhamento do fato: a tabela *Dim_Date* representa a data da venda, a tabela *Dim_Store* representa a loja que efetuou as vendas e a tabela *Dim_Product* representa o produto vendido. A característica mais marcante deste modelo que o define como do tipo estrela é a estrutura das tabelas de dimensão. Cada tabela de dimensão deste modelo está estruturada para armazenar mais de uma dimensão. A tabela *Dim_Date*, por exemplo, além de registrar as datas de venda, também armazena a descrição do dia da semana, do mês, do trimestre e do ano, que são categorizações da informação de data.

O modelo do tipo estrela possibilita uma leitura mais rápida dos dados em relação a um modelo normalizado, devido ao menor número de junções necessárias em uma consulta SQL pelo modelo estrela comumente ter um número menor de tabelas. Também por este motivo as tabelas de dimensão costumam ser muito volumosas e apresentarem grande repetição dos dados nas colunas. Usando a tabela *Dim_Date* como exemplo, a coluna *Month_Name* armazena o nome do mês correspondente à data de referência. Dessa forma, para cada dia de um determinado mês, para cada ano, o nome do mês é repetido. Em um modelo normalizado, a tabela teria somente uma coluna com uma chave estrangeira para uma tabela com os meses existentes. Sobre a questão das informações estarem armazenadas no menor grão disponível, isso possibilita uma menor necessidade de alterações no modelo pelo surgimento de uma nova necessidade da área de

negócio. No entanto, existe o risco de armazenamento de informação desnecessários, ocasionando desperdício de recursos tecnológicos.

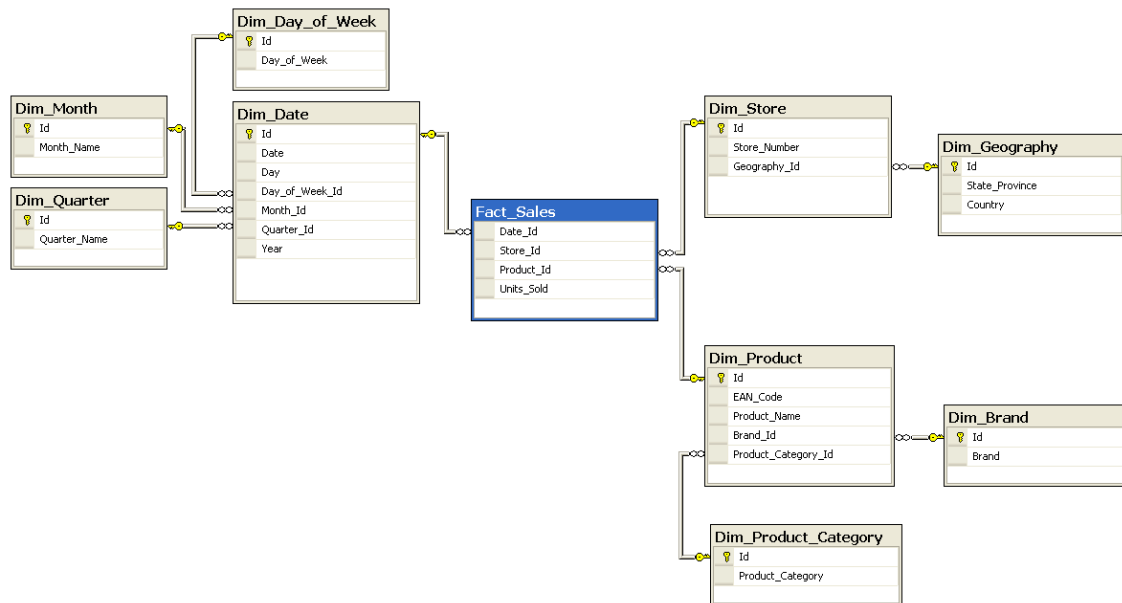


Figura 2.5: Exemplo de modelo de dados do tipo floco de neve

Para Inmon [Inmon, 2002], o *Data Warehouse* é um repositório de informações corporativas, com modelo de dados normalizado, que provê informações para carga de *Data Marts* de assuntos específicos, com maior grão. O tipo de modelagem multidimensional normalizado é chamado de *Snowflake Schema*, ou modelo floco de neve. No modelo floco de neve, cada tabela de dimensão representa uma dimensão / visão de análise e a tabela de fato apresenta o nível de detalhe necessário para análise dos fatos armazenados.

Na Figura 2.5 é ilustrado um exemplo de modelo do tipo floco de neve, também para análise de vendas de produtos. A tabela de fato *Fact_Sales* tem a mesma estrutura da Figura 2.4, armazenando o fato de unidades vendidas de determinado produto e representado pela coluna *Units_Sold*. As tabelas restantes do modelo também são tabelas de dimensão, no entanto nem todas representam um detalhamento do fato. Como o modelo de dados está normalizado, os agrupamentos das dimensões do modelo da Figura 2.5 estão armazenados como tabelas independentes. Utilizando o mesmo exemplo da tabela *Dim_Date*, neste modelo ela armazena somente as datas de venda. A descrição do dia da semana, do mês e do trimestre da data são armazenados nas tabelas *Dim_Day_of_Week*, *Dim_Month* e *Dim_Quarter*, respectivamente. A informação do ano da data da venda não tem uma tabela própria por não necessitar de descrição. A co-

luna que representaria a chave estrangeira com uma tabela de ano já é o número que representa o ano em si.

Em oposição ao modelo estrela, o modelo do tipo floco de neve normalmente apresenta um número maior de tabelas e uma leitura não tão rápida dos dados, devido ao maior número de junções necessárias em uma consulta SQL. Em contraponto, as tabelas de dimensão costumam ser menos volumosas e não apresentar repetição dos dados nas colunas. Sobre a granularidade das informações, o armazenamento no nível necessário para a análise evita o desperdício de recursos tecnológicos, ao economizar espaço de armazenamento, e processamento nos servidores pelo menor volume de dados trabalhados, mas restringe as possibilidades de análise dos usuários.

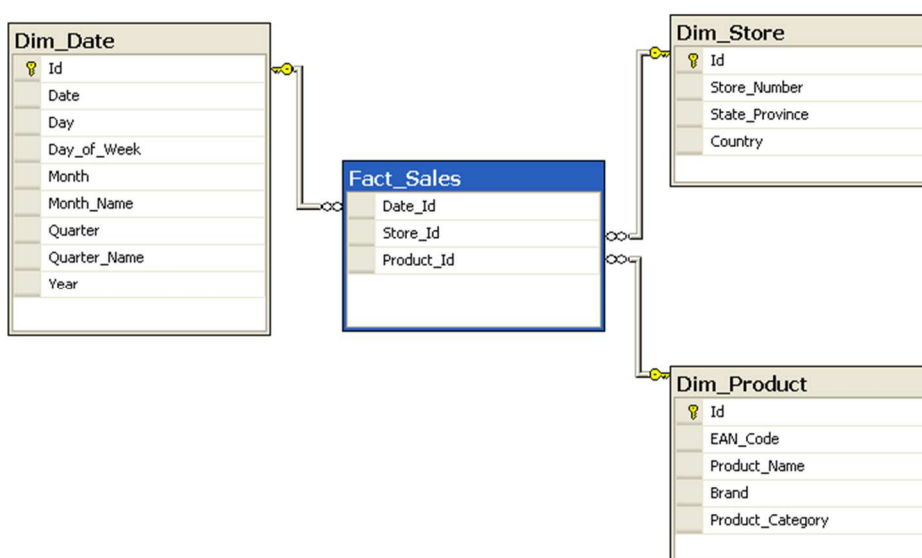


Figura 2.6: Exemplo de modelo de dados com tabela de fato sem fato

Independente do tipo de modelo de dados utilizado, estrela ou floco de neve, existe um tipo específico de tabela de fato que representa apenas o relacionamento entre tabelas de dimensão de um esquema multidimensional, sem apresentar colunas de medida. Esse tipo de tabela é chamada de tabela de Fato sem Fato (*Factless Fact*).

Na Figura 2.6 é ilustrado um exemplo de modelo de dados dimensional com uma tabela de fato sem fato, utilizando o mesmo exemplo para análise de vendas de produtos. A tabela de fato **Fact_Sales** armazena a relação entre as tabelas de dimensão de datas, dimensão de lojas e dimensão de produtos, e representa a existência de vendas para cada combinação de valores das dimensões. Como não existem colunas de medida nesta tabela fato qualificando a relação entre as tabelas de dimensão, é utilizada a contagem de registros para a realização de análises.

A arquitetura de BI ilustrada na Figura 2.1 também é complementada por servidores de camada intermediária, que disponibilizam diferentes funcionalidades, dependendo do cenário onde a aplicação de BI está inserida. Destes servidores, podemos destacar um entre os mais importantes: Servidor OLAP.

2.1.3 OLAP

OLAP (*Online Analytical Processing*) é a capacidade de analisar e manipular informação a partir de múltiplas perspectivas. Um servidor OLAP pode ser implementado mantendo uma versão multidimensional dos dados fora do banco de dados (MOLAP, ou *Multidimensional OLAP*), implementado gerando dinamicamente consultas SQL contra um banco de dados relacional (ROLAP, ou *Relational OLAP*) ou implementado em uma versão híbrida entre o MOLAP e o ROLAP (HOLAP, ou *Hybrid OLAP*). Em implementações MOLAP e HOLAP, pode-se aproveitar a disponibilidade de memória em servidores para utilização de *In-memory* BI, onde as informações em visão multidimensional são mantidas na memória do servidor para consumo [Chaudhuri, Dayal e Narasayya, 2011]. O servidor OLAP disponibiliza as informações para a camada de apresentação a partir de modelos multidimensionais existentes.

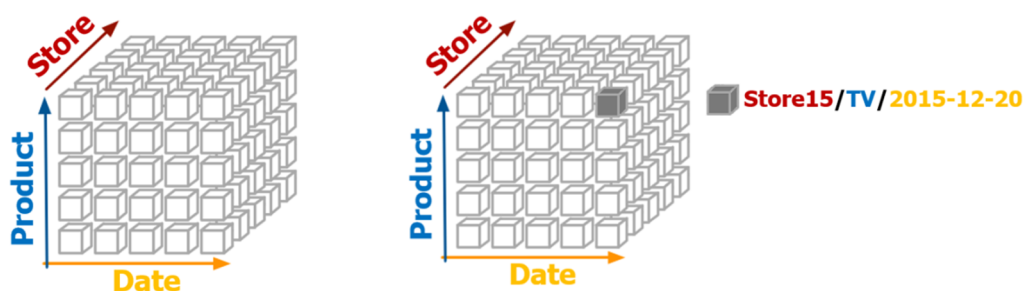


Figura 2.7: Exemplo de visão multidimensional (cubo) de dados

Na etapa de apresentação das informações, existem diversas aplicações que permitem aos usuários realizarem tarefas de BI. Dentre as mais importantes estão: a geração de planilhas e gráficos, a construção e execução de painéis interativos (*dashboards*) e a construção de relatórios sob demanda (*ad hoc*). Uma ferramenta OLAP inclui um servidor OLAP e fornece poderosa interface de usuário que permite a exploração dos dados ao longo das dimensões de análise previamente definidas [Melchert, Winter and Klesse, 2004] e que os usuários façam análises complexas através de um acesso rápido e interativo de informações, a partir de alguns pontos de vista [Airinei e Homocianu, 2009].

Na Figura 2.7 é ilustrada uma visão multidimensional dos dados, também chamada de cubo de dados, para o exemplo utilizado anteriormente de análise de vendas de produtos de uma rede de lojas. Cada dimensão do modelo de dados é representada por uma dimensão do cubo e cada célula que forma o cubo representa o cruzamento de um elemento de cada dimensão com um valor de medida associado a eles. A célula destacada armazena as unidades vendidas para a loja Store15, para o produto TV na data de 20/12/2015.

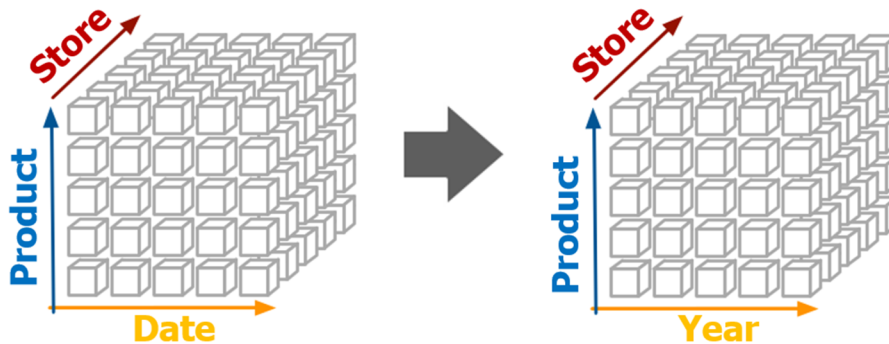


Figura 2.8: Exemplo de operação de *roll up*

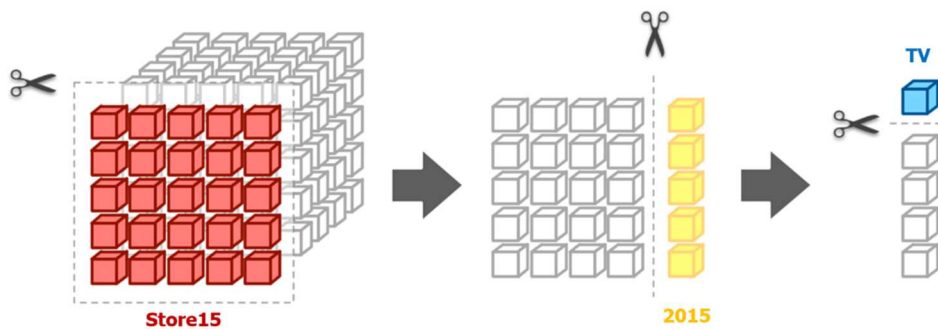


Figura 2.9: Exemplo de operação de *dice*

As ferramentas OLAP proporcionam aos seus usuários a capacidade de manipular os dados através da execução de operações analíticas definidas como operações OLAP. As operações OLAP típicas são:

- *Roll up*: representa o aumento da granularidade de uma informação, tornando-a mais consolidada, e apresenta uma granularidade de origem (antes da operação) e outra de destino (após a operação). Normalmente é realizada nas ferramentas OLAP com a substituição de uma dimensão em uma análise ou relatório por outra relacionada a ela, mas também é possível de ser feita com a retirada de uma dimensão. Na Figura 2.8 é apresentada uma operação de *roll up* sobre o cubo de dados de exemplo da Figura 2.7. O cubo de dados original teve a dimensão Date

substituída pela dimensão *Year*, armazenando as vendas por ano, ao invés de armazená-las por dia;

- *Drill down*: é a operação inversa do *roll up* e representa a diminuição da granularidade de uma informação, tornando-a mais detalhada. Normalmente é realizada nas ferramentas OLAP com a substituição de uma dimensão em uma análise ou relatório por outra relacionada a ela, mas também é possível de ser feita com a inclusão de uma dimensão;
- *Slice*: é a operação de filtro em uma das dimensões existentes no cubo de dados. Na Figura 2.9 são apresentados 3 exemplos desta operação, para a 3 dimensões existentes, com a realização de “cortes” transversais no cubo. Na primeira operação da Figura 2.9, por exemplo, são filtradas as vendas para a loja Store15;
- *Dice*: é a operação de filtro em mais de uma das dimensões existentes no cubo de dados de forma sequencial. Na Figura 2.9 é apresentado um exemplo desta operação, resultando em uma versão menor do cubo com apenas uma célula, que armazena as unidades vendidas para a loja Store15, para o produto TV no ano de 2015;
- *Pivot*: é a operação de troca de posição de uma dimensão em um relatório tabular (de linha para coluna, ou vice-versa) ou em um gráfico (como a mudança de eixo num gráfico de linha, por exemplo);

Em cenários típicos de aplicação de BI, é disponibilizada uma ferramenta OLAP para os usuários realizarem análises sobre os dados. Uma representação do conhecimento sobre essa aplicação possibilitaria a um analista de negócio ou um analista de BI conhecer as medidas disponíveis para consulta, por quais visões de análise estão disponíveis (granularidade/agregabilidade) e as relações entre essas visões (hierarquias), por exemplo, para a criação de relatórios e realizar operações OLAP.

2.2 Ontologia

Uma ontologia é uma especificação formal e explícita de uma conceituação compartilhada, sendo amplamente utilizada para uma representação formal mais rica de conhecimento interpretável por máquina [Gruber, 1993]. Nesta definição, “conceituação” se refere a um modelo abstrato de algum conhecimento de domínio que identifica os conceitos relevantes. “Compartilhada” indica que uma ontologia captura o conheci-

mento consensual, ou seja, compartilhado por um grupo. “Explícita” significa que a representação da ontologia, incluindo conceitos, relacionamentos e as restrições sobre estes conceitos, são definidos explicitamente, de forma acessível a outros. Finalmente, “formal” significa que a ontologia deve ser interpretável por máquina.

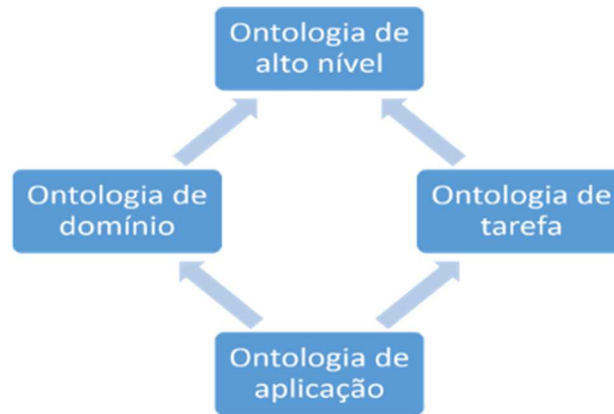


Figura 2.10: Classificação de ontologias

Segundo Guarino [Guarino, 1997], podemos classificar ontologias como de alto nível, de domínio, de tarefa ou de aplicação. Essas classificações de ontologia se relacionam de forma hierárquica, conforme mostrado na Figura 2.10, sendo a ontologia de alto nível mais abrangente e a ontologia de aplicação mais específica.

As ontologias de alto nível descrevem conceitos mais genéricos, como conceitos relativos a tempo e espaço, e independentes de um problema ou assunto particular. As ontologias de domínio especializam conceitos de alto nível para abordar conceitos e descrever o vocabulário de um domínio ou assunto mais específico, como esporte e informática por exemplo. De forma semelhante, as ontologias de tarefa especializam conceitos de alto nível para abordar conceitos e descrever o vocabulário de uma tarefa ou atividade mais específica, como cozinhar e pilotar por exemplo. As ontologias de aplicação especializam conceitos de domínio e tarefa e descrevem o vocabulário de uma aplicação, como pilotagem de aviões por exemplo.

A estrutura de uma ontologia é geralmente formada por:

- Classes, que representam os conceitos de um domínio, tarefa ou aplicação;
- Instâncias, que representam os objetos ou indivíduos particulares;
- Relações, que representam as associações entre os conceitos.

Adicionalmente, uma ontologia pode possuir também axiomas, que são sentenças / afirmações que descrevem regras formais de um domínio de aplicação e restringem os valores para classes ou instâncias e as possíveis interpretações dos termos.

Uma definição mais concisa de uma ontologia é uma tupla $\langle C, R, I, A \rangle$, onde “C” é um conjunto de conceitos, “R” é um conjunto de relações, “I” é um conjunto de instâncias e “A” um conjunto de axiomas [Staab and Studer, 2004]. Ontologias que possuem axiomas e um mecanismo de raciocínio construído sobre eles são chamadas de ontologias formais [Mansingh, Osei-Bryson and Reichgelt, 2011].

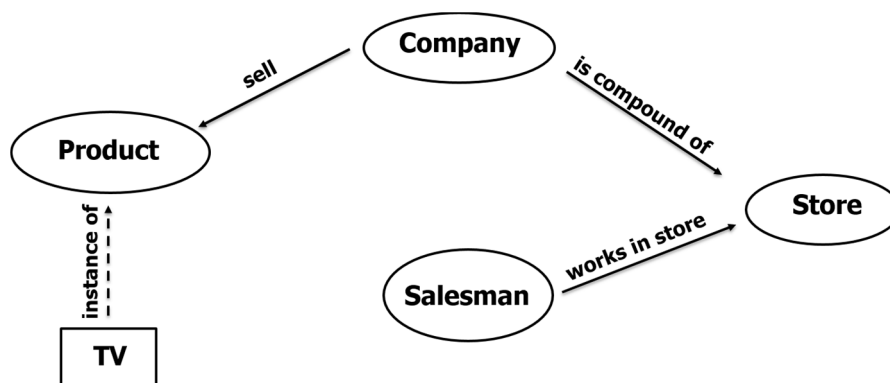


Figura 2.11: Exemplo de ontologia de domínio

Na Figura 2.11 é ilustrado um exemplo de ontologia, que é uma possível representação do domínio de uma rede de lojas de varejo, para vendas de produtos a clientes. Nesta ontologia, os conceitos estão representados como elipses, como **Product** por exemplo, as relações entre os conceitos estão representadas como setas, como **sell** por exemplo, representando que a companhia vende produtos, e as instâncias como retângulos, como o produto **TV**. Um possível exemplo de axioma para esta ontologia poderia ser que cada companhia deve ter pelo menos uma loja.

A tarefa associada com a construção de ontologias inclui a extração de conceitos relevantes, construindo hierarquias “*is a*” ou “é um” (relações entre classes), e extraindo e formalmente definindo os relacionamentos entre os conceitos. Para permitir raciocínio, axiomas devem ser representados em uma linguagem lógica bem compreendida [Mansingh, Osei-Bryson and Reichgelt, 2011].

2.2.1 OWL

Ontologias podem ser representadas utilizando diferentes formalismos de representação de conhecimento. Existem uma grande variedade de formalismos para expressar ontologias [Staab and Studer, 2013]. Dentre os formalismos para especificar ontologias, a *Ontology Web Language* (OWL) é o mais usual. A OWL é um padrão definido pelo *World Wide Web Consortium* (W3C), consórcio internacional que é a principal organização de padronização da *World Wide Web* e tem a finalidade de estabelecer padrões para a criação e a interpretação de conteúdos para a Web.

A OWL descreve classes, propriedades e relações entre esses objetos conceituais em uma forma que facilita a interpretabilidade de máquina de conteúdo Web. A linguagem OWL possui três sub-linguagens, dispostas em uma sequência crescente de níveis de expressividade [Breitman, Casanova and Truszkowski, 2007]:

- OWL-Lite: oferece hierarquias de classes, com suas propriedades, e restrições com expressividade suficiente para modelar ontologias simples. OWL-Lite impõe limitações em como as classes se relacionam entre si [W3C, 2016]. A sub-linguagem OWL-Lite tem restrições de expressividade, como somente suportar cardinalidade 0 ou 1 para as constraints, por exemplo;
- OWL-DL: utiliza o vocabulário completo de OWL sem perder a completude computacional (todos os vínculos são computáveis) e decidibilidade (todos os cálculos terminam em tempo finito) de sistemas de raciocínio. OWL-DL inclui todas as construções de linguagem OWL com restrições, tais como a separação tipo (uma classe não pode ser também um indivíduo). OWL-DL deriva da Lógica Descritiva existente e tem propriedades computacionais desejáveis para sistemas de raciocínio [W3C, 2016];
- OWL-Full: é destinado a usuários que querem máxima expressividade sem garantias computacionais. Por exemplo, em OWL-Full uma classe pode ser tratada simultaneamente como uma coleção de indivíduos e como um indivíduo. OWL-Full permite a uma ontologia aumentar o significado do vocabulário pré-definido. É improvável que qualquer software de raciocínio seja capaz de suportar todas as funcionalidades do OWL-Full [W3C, 2016]. A sublinguagem OWL-Full apresenta máxima expressividade, mas não oferece garantia computacional.

Uma classe pode ser tratada simultaneamente como um indivíduo e como uma coleção de indivíduos, por exemplo.

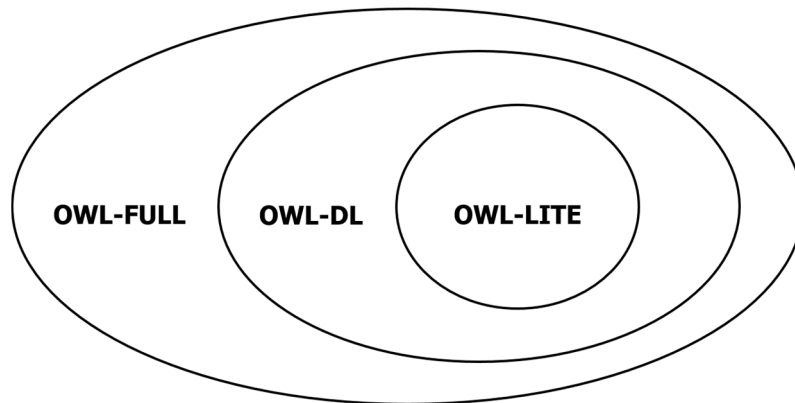


Figura 2.12: Relação entre sub-linguagens OWL

Cada uma destas sub-linguagens é uma extensão de sua predecessora, tanto em relação ao que pode ser expressado, como em relação ao que pode ser derivado [W3C, 2016], conforme ilustrado na Figura 2.12.

2.3 Considerações finais

A característica distinta de ontologias é a existência de uma semântica teórica de modelo: ontologias são teorias lógicas. A interpretação de ontologia não é deixada para os usuários que lêem os diagramas ou aos sistemas de gestão do conhecimento que as implementam, ela é especificada explicitamente. A semântica fornece as regras para interpretar a sintaxe que não fornecem o significado diretamente, mas restringe as possíveis interpretações do que é declarado [Euzenat and Shvaiko, 2013].

Em cenários típicos de sistemas de BI, uma representação de conhecimento permitiria a um analista de negócios ou a um analista de BI conhecer as medidas disponíveis para análise, por quais visões de análise estão disponíveis (granularidade / agregabilidade) e as relações entre essas visões, por exemplo.

Assim sendo, seria de grande utilidade uma ontologia de aplicação descrevendo os conceitos implementados no *Data Warehouse*, de forma a apoiar sua utilização.

3 – Proposta de Solução

Este capítulo tem como propósito descrever a proposta desta pesquisa para derivação automática de ontologias de aplicação de sistemas de BI a partir de Data Warehouses e a ferramenta desenvolvida que implementa a proposta.

Para exemplificar a solução proposta neste trabalho, considere um cenário real de uma aplicação de BI sobre os funcionários de uma instituição financeira participantes de um fundo de pensão.

Aos analistas de negócio é disponibilizada uma ferramenta OLAP para a realização de análises e criação de relatórios e painéis interativos (*dashboards*). Essas informações são utilizadas para diversas funções na empresa. No entanto, os analistas de negócio são muito dependentes de especialistas de TI para a construção de novos relatórios e realização de análises, devido à dificuldade em conhecer as informações que existem disponíveis para uso e as possibilidades de cruzamento entre elas. Aliado a isso, com o impulso de funcionalidades de cruzamento de informações nas ferramentas OLAP, a demanda por conhecimento sobre os dados disponíveis tende a ser maior. Entretanto, a área de TI não tem disponibilidade suficiente para atender a demanda e a documentação disponível não acompanha as mudanças que ocorrem no ambiente tecnológico.

Nesse cenário, uma descrição dos conceitos e seus relacionamentos disponíveis para análise no ambiente de BI forneceria aos usuários conhecimento que lhes permite tomar melhores decisões. Além disso, ela também poderia auxiliar os analistas de BI ao explicitar discrepâncias entre o esquema de dados e a camada de aplicação e ao apoiar demandas de integração de dados. Como exemplo, imaginemos uma situação onde uma área de negócio deste fundo de pensão necessita realizar uma análise sobre o impacto financeiro da aposentadoria de funcionários da instituição financeira. O cálculo do valor da aposentadoria de um funcionário se baseia, dentre outras variáveis, no valor de seu salário. Na ferramenta OLAP disponibilizada um analista de negócio visualiza facilmente uma métrica com o salário do funcionário. Entretanto, outras informações como

as possibilidades de agrupamento ou filtragem dessa informação, as dimensões possíveis de serem utilizadas ou a granularidade da informação não são disponibilizadas. Além disso, outras informações relacionadas àquela escolhida poderiam prover mais *insights* para uma melhor decisão, tais como a quantidade de dependentes, o sexo e a cidade de residência. Uma ontologia desta aplicação poderia representar Salário e Quantidade de dependentes como métricas associadas a uma dimensão temporal e a outras dimensões relativas aos funcionários, como Sexo e Município de residência. Então, esta ontologia poderia ser utilizada como um artefato para prover conhecimento extra para uma melhor tomada de decisão.

Assim, uma ontologia contendo os conceitos implementados no DW seria de grande utilidade, tanto para analistas de negócio quanto para analistas de BI, como uma representação atualizável do que está disponível no banco de dados. Os analistas de negócio a utilizariam como apoio/complemento à ferramentas OLAP, maximizando seu uso e evitando a realização de análises incorretas. Os analistas de BI podem utilizá-la para apontar discrepâncias ou implementações incorretas no esquema de dados.

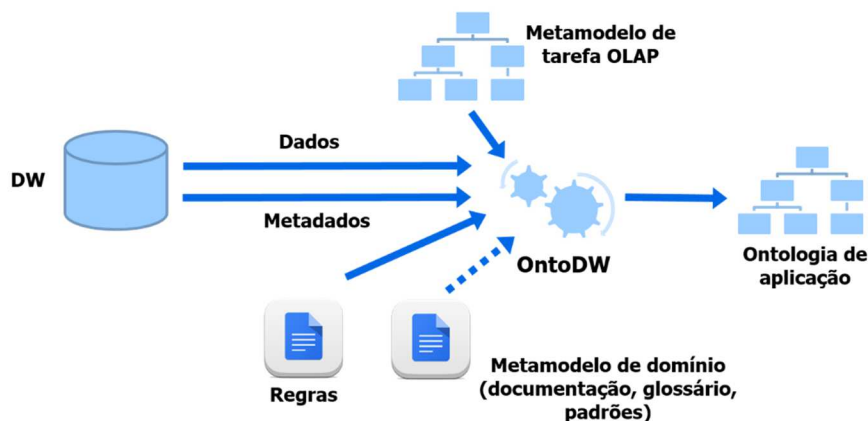


Figura 3.1: Visão geral da geração de ontologia

Para preencher esta lacuna, este trabalho propõe OntoDW, uma abordagem para a extração automática de uma ontologia a partir dos construtos estruturais (metadados do esquema) e conteúdo (dados) de *Data Warehouses* dentro de um sistema de BI. Os elementos da ontologia gerada são obtidos através de regras de mapeamento específicas, que compõem a abordagem proposta. Neste trabalho é também descrita a ferramenta desenvolvida com a abordagem proposta, de forma a se obter uma ontologia com os conceitos mapeados a partir de um DW.

A hipótese para essa proposta é que é possível gerar ontologias de aplicação a partir de *Data Warehouses* através do uso de regras de mapeamento específicas, e essas ontologias refletirão o conhecimento relativo à tarefa de análise OLAP presente nos dados e metadados dos *Data Warehouses*. Os metadados de ferramentas OLAP ou de ETL não foram considerados pois foi definido como premissa que nem sempre estão disponíveis, seja pelo não uso na solução de BI ou por restrições do fabricante às suas estruturas de dados. A ontologia obtida deve contemplar não somente o conhecimento já explícito nas estruturas de dados (como traduzir tabelas para classes, por exemplo), mas também a semântica que está implícita (como categorizações de classes, por exemplo).

Uma ontologia de domínio compreenderia os conceitos do negócio que estão estruturados no *Data Warehouse*, sem especificar as possibilidades de operação para realizar. Por esta razão, esta solução proposta gera uma ontologia de aplicação e inclui classes relacionadas a um metamodelo de tarefa OLAP.

A ontologia gerada será composta de conceitos de negócio refletidos no esquema de dados multidimensional (tabelas de fato e tabelas de dimensão) e conceitos associados às operações analíticas em sistemas de BI (como agregabilidades, as análises possíveis de se realizar). Para isso, são utilizados como elementos de entrada para o processo de geração da ontologia: o DW, um metamodelo de tarefa OLAP, um metamodelo de domínio e o conjunto de regras de mapeamento definidos nesta proposta, como na Figura 3.1. A execução destas regras em sequência funciona como um processo para a geração dos conceitos, de forma automática, que farão parte da ontologia da aplicação gerada.

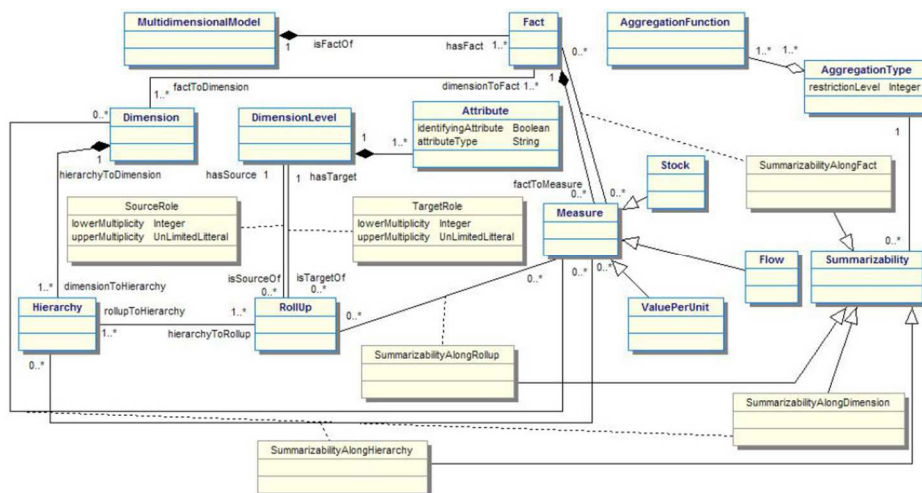


Figura 3.2: Metamodelo de tarefa OLAP [Prat, Megdiche and Akoka, 2012]

A ontologia de aplicação será gerada em formato OWL, por ser a linguagem recomendada pela W3C para representação de conhecimento, por permitir expressar conceitos e a relação entre eles e pela sua facilidade de utilização. A sublinguagem utilizada para a ontologia será a OWL-DL, por permitir uma expressividade adequada da linguagem OWL em um artefato computável. OntoDW objetiva a geração de uma ontologia com os conceitos e seus relacionamentos identificados no DW. Entretanto, é esperado que a ontologia seja computável para permitir sua utilização para outras finalidades diferentes das previstas neste trabalho.

O metamodelo de tarefa OLAP apresenta as classes e relacionamentos predefinidos associados à análise de informações em sistemas de BI, como as classes **Measure**, **RollUp** e **Dimension**, por exemplo. No presente trabalho, foi adotado o metamodelo de tarefa OLAP proposto por Prat et al. [Prat, Megdiche and Akoka, 2012], ilustrado na Figura 3.2, que estrutura os conceitos de um ambiente de BI (conceitos de modelos dimensionais e de aplicações OLAP), e como tais conceitos se inter-relacionam. Este metamodelo reflete as principais tarefas que compõem a teoria sobre operações analíticas e foi utilizado em trabalhos aceitos em eventos de relevância.

O metamodelo do domínio apresenta conceitos específicos do domínio no qual a aplicação está inserida e pode ser representado por um dicionário de dados, um padrão de terminologia ou um glossário, que são simples componentes tradicionalmente encontrados em ambientes organizacionais. As informações do metamodelo do domínio serão utilizadas para nomear conceitos de acordo com termos de negócio já estabelecidos.

Para obter uma ontologia mais rica de conceitos e mais alinhada a uma aplicação de BI, a presente proposta definiu regras de mapeamento (descritas a seguir) para as seguintes classes do metamodelo de tarefa OLAP de Prat et al. [Prat, Megdiche and Akoka, 2012]:

- **Fact**: Classe que representa a tabela de fato do esquema de dados multidimensional. Pode também ser interpretada como um agrupamento de medidas.
- **Dimension**: Classe que representa a tabela de dimensão do esquema de dados multidimensional. Pode também ser interpretada como uma visão de análise dos dados, podendo ter 1 ou mais níveis.

- **DimensionLevel:** Classe que representa o nível de uma dimensão, ou subdivisão de uma visão de análise. Uma dimensão pode apresentar mais de um nível se a tabela de onde ela foi mapeada estiver desnormalizada.
- **Attribute:** Classe que representa uma propriedade / qualificador de um nível de dimensão. Pode ser identificador ou não no nível de dimensão.
- **Measure:** Classe que representa o indicador que se quer analisar.
- **RollUp:** Classe que representa uma operação possível de mudança de granularidade / visão de análise de uma medida. Cada instância é composta de dois níveis de dimensão, que representam a granularidade de origem e a granularidade de destino.
- **Hierarchy:** Classe que representa um conjunto de rollups relacionados. Pode representar um agrupamento de rollups entre níveis de uma mesma dimensão ou entre níveis de diferentes dimensões.
- **SummarizabilityAlongFact:** Classe que representa a agregabilidade de uma medida ao longo de fatos. Ou seja, por quais fatos a medida se relaciona, em quais fatos está presente.
- **SummarizabilityAlongDimension:** Classe que representa a agregabilidade de uma medida ao longo de dimensões. Ou seja, por quais dimensões a medida pode ser analisada.
- **SummarizabilityAlongHierarchy:** Classe que representa a agregabilidade de uma medida ao longo de hierarquias. Ou seja, por quais hierarquias a medida pode ser analisada.

Estes conceitos não representam a totalidade das classes presentes no metamodelo de tarefa utilizado, mas são os principais conceitos para uma ontologia rica e alinhada para o sistema de BI. As classes do metamodelo de tarefa OLAP de Prat et al. [Prat, Megdiche and Akoka, 2012] não utilizadas para este trabalho, juntamente com as respectivas justificativas por não considerá-las, foram as seguintes:

- **MultidimensionalModel:** Classe que representa um modelo dimensional como um todo. Não será utilizada, pois cada ontologia gerada representará apenas um esquema de dados.
- **AggregationFunction:** Classe que representa as possíveis funções de agregação utilizadas para consolidação do valor de uma medida ao ser realizada uma ope-

ração de aumento da granularidade. Não utilizada por não estar alinhada com o foco deste trabalho, que visa explicitar os conceitos para utilização na construção de relatórios e as análises possíveis de serem realizadas.

- **AggregationType:** Classe que representa os tipos possíveis de agregação de uma medida, que pode ser aditiva, semi-aditiva ou não aditiva. Não utilizada por não estar alinhada com o foco deste trabalho, que visa explicitar os conceitos para utilização na construção de relatórios e as análises possíveis de serem realizadas.
- **Stock:** Classe que representa as medidas qualificadas como do tipo estoque. Não utilizada por não estar alinhada com o foco deste trabalho, que visa explicitar os conceitos para utilização na construção de relatórios e as análises possíveis de serem realizadas.
- **Flow:** Classe que representa as medidas qualificadas como do tipo fluxo. Não utilizada por não estar alinhada com o foco deste trabalho, que visa explicitar os conceitos para utilização na construção de relatórios e as análises possíveis de serem realizadas.
- **ValuePerUnit:** Classe que representa as medidas qualificadas como do tipo valor por unidade. Não utilizada por não estar alinhada com o foco deste trabalho, que visa explicitar os conceitos para utilização na construção de relatórios e as análises possíveis de serem realizadas.
- **SourceRole:** Classe que representa a cardinalidade do nível de dimensão de origem de um rollup. Não utilizada por não estar alinhada com o foco deste trabalho, que visa explicitar os conceitos para utilização na construção de relatórios e as análises possíveis de serem realizadas, e por representar uma informação demasiadamente detalhada.
- **TargetRole:** Classe que representa a cardinalidade do nível de dimensão de destino de um rollup. Não utilizada por não estar alinhada com o foco deste trabalho, que visa explicitar os conceitos para utilização na construção de relatórios e as análises possíveis de serem realizadas, e por representar uma informação demasiadamente detalhada.

- **Summarizability:** Classe que representa a agregabilidade de uma medida, ou possibilidades de análise e mudança de granularidade. Não utilizada por ser apenas um supertipo dos diferentes tipos de classe de agregabilidade utilizadas.
- **SummarizabilityAlongRollup:** Classe que representa a agregabilidade de uma medida ao longo de rollups. Não utilizada por ser uma informação já disponibilizada ao se definir a agregabilidade ao longo de hierarquias. Os rollups que compõem uma hierarquia estarão apresentados na ontologia gerada.

3.1 Regras de Mapeamento

Esta Seção descreve as regras de mapeamento definidas entre os elementos de *Data Warehouses* (dados e metadados) e os conceitos da ontologia do metamodelo de tarefa OLAP utilizado (Figura 3.2).

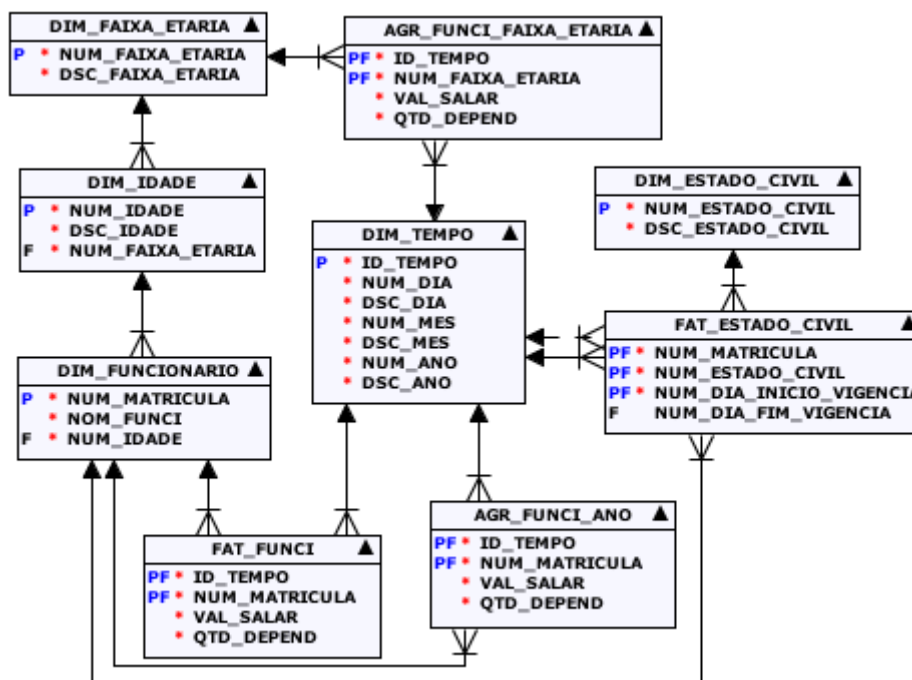


Figura 3.3: Diagrama do esquema de dados multidimensional do domínio de cadastro de funcionários

As regras de mapeamento são descritas ilustrando exemplos da sua aplicação no esquema de dados multidimensional da Figura 3.3, elaborado a partir da adaptação e simplificação de um *Data Warehouse* real utilizado em um fundo de pensão brasileiro para análise do seu corpo social, ou seja, dos funcionários que podem ser e dos que já

são seus participantes. Os funcionários são classificados de acordo com o estado civil e idade (ou faixa etária). As informações dos funcionários incluem 2 medidas: o salário do funcionário (em reais por mês) e quantidade de dependentes do funcionário.

As regras de mapeamento definidas contemplam esquemas de dados que implementam técnicas diversas de modelagem multidimensional. As regras tratam desde estruturas mais comumente encontradas como tabelas de fato e tabelas de dimensão normalizadas, até estruturas de análise mais complexa, como tabelas de agregação, tabelas de fato sem fato e dimensões desnormalizadas.

Para as regras definidas nas seções a seguir, considere E um esquema lógico relacional de um DW genérico, e seja $T = \{T_1, T_2, \dots, T_n\}$ o conjunto de tabelas (ou relações) pertencentes ao esquema E.

3.1.1 Classe Fact

Assume-se que existe um fato (ou uma tabela de fatos no esquema do DW) para cada tabela que tenha ao menos uma coluna como chave estrangeira, mas que não seja referenciada por nenhuma chave estrangeira de outra tabela do esquema do DW; ou seja, mapeia-se um fato F para cada tabela T1 que contenha uma coluna C como chave estrangeira para T2, se a chave primária PK de T1 não for alvo de nenhuma chave estrangeira de qualquer tabela T3. Esta regra se justifica pela própria definição do esquema multidimensional. Assume-se que F é um Fato sem Fato se não houver colunas de tipo numérico fora da chave primária, que também não possuam chave estrangeira.

Formalmente, os conceitos **Fact** são extraídos aplicando-se a regra R1 definida na Figura 3.4:

| |
|--|
| <p>Regra R1: Para cada tabela T1, T1 é mapeada para um fato F1 se e somente se não houver nenhuma tabela T2 ($T_2 \neq T_1$) que referencie T1 através de chave estrangeira e T1 referencie uma tabela T3 através de chave estrangeira. Seja $PK = \{C_1, \dots, C_i\}$ o subconjunto de colunas de F1 que compõem a sua chave primária e $NK = \{C_{i+1}, \dots, C_n\}$ o subconjunto de colunas de F1 que não fazem parte da sua chave primária. Para cada fato F1, F1 é qualificada como Fato sem Fato se não tiver nenhuma coluna X ($X \in NK$) de tipo numérico como chave estrangeira.</p> |
|--|

Figura 3.4: Regra R1 - Mapeamento de conceitos Fact

Exemplos de conceitos **Fact** extraídos aplicando-se a regra R1 sobre o esquema da Figura 3.3 são:

- FAT_FUNCI;
- AGR_FUNCI_ANO;
- AGR_FUNCI_FAIXA_ETARIA;
- FAT_ESTADO_CIVIL.

3.1.2 Classe Dimension

Assume-se que existe uma dimensão (ou uma tabela de Dimensão no esquema do DW) para cada tabela que seja referenciada por alguma chave estrangeira de outra tabela do esquema do DW; ou seja, mapeia-se uma dimensão D para cada tabela T1 que a chave primária PK de T1 for alvo de alguma chave estrangeira de qualquer tabela T2. Esta regra se justifica pela própria definição do esquema multidimensional. Adicionalmente:

- Assume-se que D é uma Dimensão de Tempo se houver colunas de tipo data populada com datas sequenciais sem intervalos e sem valores repetidos (comumente, todo DW apresenta pelo menos uma dimensão temporal, caso contrário as análises não ficam contextualizadas e não permitem observar evoluções);
- Assume-se que D é uma Dimensão Versionada (ou Dimensão de Modificação Lenta) se houver colunas de tipo data que não estejam populadas com datas sequenciais sem intervalos e sem valores repetidos. Uma Dimensão Versionada é aquela que determina período de vigência para cada valor;
- Caso não seja uma Dimensão de Tempo ou uma Dimensão Versionada, D é uma Dimensão de Negócio.

Formalmente, os conceitos **Dimension** são extraídos aplicando-se a regra R2 definida na Figura 3.5:

| |
|--|
| <p>Regra R2: Para cada tabela T1, T1 é mapeada para uma dimensão D1 se houver alguma tabela T2 (T2 ≠ T1) que referencie T1 através de chave estrangeira. Seja CL={C1,...,Cn} o conjunto de colunas de D1 e VCm={VCm1,...,VCmn} o conjunto de valores inseridos em uma coluna Cm. Adicionalmente:</p> <ul style="list-style-type: none">• D1 é qualificada como Dimensão de Tempo se tiver coluna X (X ∈ CL) de tipo data e $\forall a, VX_{a+1} = VX_a + 1$; |
|--|

- D1 é qualificada como Dimensão Versionada se tiver coluna X ($X \in CL$) de tipo data e $\forall a, VX_{a+1} \neq VX_a + 1$;
- Caso não seja uma Dimensão de Tempo ou Dimensão Versionada, D1 é qualificada como Dimensão de Negócio.

Figura 3.5: Regra R2 - Mapeamento de conceitos **Dimension**

Exemplos de conceitos **Dimension** extraídos aplicando-se a regra R2 sobre o esquema da Figura 3.3 são:

- DIM_FUNCIONARIO;
- DIM_IDADE;
- DIM_FAIXA_ETARIA;
- DIM_ESTADO_CIVIL;
- DIM_TEMPO.

3.1.3 Classe DimensionLevel

Assume-se que existe um nível de dimensão para cada conjunto de colunas de uma dimensão onde nenhuma coluna esteja como chave estrangeira e se, para cada valor de uma coluna qualquer desse conjunto, ocorrer sempre o mesmo valor correspondente em outra coluna deste mesmo conjunto. As colunas que estejam como chave estrangeira não são consideradas, pois apenas apresentam a relação com outro nível. A questão dos valores das colunas se justifica pelo fato de que um registro num nível de dimensão deve ser único, como se fosse uma outra tabela de dimensão.

Formalmente, os conceitos **DimensionLevel** são extraídos aplicando-se a regra R3 definida na Figura 3.6:

Regra R3: Para cada dimensão D1, seja $ND=\{C_1, \dots, C_j\}$ um subconjunto de colunas de D1, ND é mapeado para um nível de dimensão N1 se, $\forall a, Ca$ ($Ca \in ND$) não possui chave estrangeira e, $\forall b$, um valor qualquer em Ca ($Ca \in ND$) sempre tem um mesmo valor correspondente na coluna Cb ($Cb \in ND$). Adicionalmente:

- N1 recebe o mesmo nome de D1 se for o único nível de D1 ou se, caso não seja o único nível de D1, N1 apresentar a mesma cardinalidade de D1; ou seja, as colunas de N1 apresentarem a mesma quantidade de registros distintos que as colunas de D1;
- Caso não receba o mesmo nome de D1, N1 recebe um nome baseado no nome de seu atributo identificador.

Figura 3.6: Regra R3 - Mapeamento de conceitos **DimensionLevel**

Como exemplo, para a dimensão DIM_TEMPO mapeada pela regra R2, um grupo de colunas mapeado para nível de dimensão é $ND = \{ID_TEMPO, NUM_DIA, DSC_DIA\}$, pois as colunas não tem chave estrangeira e a cardinalidade entre quaisquer duas colunas é sempre de 1 para 1. Na dimensão DIM_TEMPO cada registro representa um dia do ano, exemplificado na Tabela 3.1. Referenciando o registro na Tabela 3.1, o valor “20151210” na coluna NUM_DIA sempre estará associado ao valor “237625” da coluna ID_TEMPO e ao valor “10/12/2015” da coluna DSC_DIA. Como este nível representa a mesma granularidade da dimensão, seu nome será DIM_TEMPO.

Tabela 3.1: Exemplo de registro da tabela DIM_TEMPO

| Coluna | ID_TEMPO | NUM_DIA | DSC_DIA | NUM_MES | DSC_MES | NUM_ANO | DSC_ANO |
|--------|----------|----------|------------|---------|---------|---------|---------|
| Valor | 237625 | 20151210 | 10/12/2015 | 201512 | 12/2015 | 2015 | 2015 |

Outros exemplos de conceitos **DimensionLevel** extraídos aplicando-se a regra R3 sobre o esquema da Figura 3.3 são:

- DIM_FUNCIONARIO;
- DIM_IDADE;
- DIM_FAIXA_ETARIA;
- DIM_ESTADO_CIVIL;
- MES (dimensão DIM_TEMPO, colunas NUM_MES e DSC_MES);
- ANO (dimensão DIM_TEMPO, colunas NUM_ANO e DSC_ANO).

3.1.4 Classe Attribute

Toda coluna de tabela no DW que pertença a um nível de dimensão N1 será um atributo A1. A1 será um atributo identificador caso a coluna que o originou componha a chave primária da tabela onde está inserida, ou caso a coluna que o originou não componha a chave primária, mas a coluna for de tipo numérico e N1 não tiver outra coluna pertencente à chave primária. Esta última definição se baseia em convenção normalmente utilizada de definir colunas identificadoras com tipo de dado numérico, seja incrementada através de *sequences* ou com a inserção dos registros de forma não automática.

Formalmente, os conceitos **Attribute** são extraídos aplicando-se a regra R4 definida na Figura 3.7:

| |
|--|
| <p>Regra R4: Para cada nível de dimensão $N1$, seja $ND=\{C1, \dots, Cj\}$ o conjunto de colunas de $N1$, $PK=\{Cm, \dots, Cn\}$ o conjunto de colunas que integram a chave primária da tabela que originou $N1$; para cada Ca ($Ca \in ND$), Ca é mapeado para um atributo $A1$. Para cada atributo $A1$, $A1$ é qualificado como atributo identificador (id) se $Ca \in PK$, ou se $Ca \notin PK$, Ca é de tipo numérico e $ND \cap PK = \emptyset$.</p> |
|--|

Figura 3.7: Regra R4 - Mapeamento de conceitos **Attribute**

Como exemplo, para o nível de dimensão DIM_TEMPO, mapeado pela regra R3 e exemplificado na subseção anterior, todas as colunas foram mapeadas em atributos. O atributo ID_TEMPO foi qualificado como identificador, pois a coluna ID_TEMPO pertence à chave primária da tabela de dimensão DIM_TEMPO.

Outros exemplos de conceitos **Attribute** foram extraídos aplicando-se a regra R4 sobre o esquema da Figura 3.3. Nos exemplos, após o nome do atributo, são apresentados entre parênteses o nível de dimensão que o contém e a indicação de atributo indicador, se for o caso:

- NUM_MATRICULA (DIM_FUNCIONARIO) (atributo identificador);
- NOM_FUNCI (DIM_FUNCIONARIO);
- NUM_MES (MES) (atributo identificador);
- DSC_MES (MES).

3.1.5 Classe Measure

Existem 2 cenários para o mapeamento de medidas, porque as medidas pertencentes as tabelas de fato sem fato apresentam medidas implícitas, enquanto as tabelas de fato restantes apresentam medidas explícitas, através de colunas. No cenário 1, assume-se que em um fato F1, as colunas fora da chave primária que não tenham chave estrangeira e que sejam de tipo numérico são medidas. A coluna deve ser numérica para possibilitar operações sobre o valor presente nela, como soma ou média, por exemplo. O nome da medida será o mesmo nome da coluna. No cenário 2, caso F1 seja um Fato sem Fato, existirá uma medida mea-F1 sem coluna correspondente na tabela do DW. O nome

da medida recebe o nome da tabela que a contém após o prefixo “mea-”, para que seja reproduzida a semântica embutida no nome da tabela.

Formalmente, os conceitos **Measure** são extraídos aplicando-se as regras R5 e R6 definidas nas Figuras 3.8 e 3.9, respectivamente:

Regra R5: Para cada fato F1, seja $NK=\{C_i, \dots, C_n\}$ o subconjunto de colunas de F1 que não fazem parte da sua chave primária. Para cada C_a ($C_a \in NK$), C_a é mapeada para uma medida M1 se for uma coluna de tipo numérico e não possuir uma chave estrangeira.

Figura 3.8: Regra R5 - Mapeamento de conceitos **Measure**

Regra R6: Para cada fato F1 classificado como Fato sem Fato, F1 é mapeado para uma medida M1.

Figura 3.9: Regra R6 - Mapeamento de conceitos **Measure**

Como exemplo, para a tabela de fato FAT_FUNCI mapeada pela regra R1 sobre o esquema da Figura 3.3, o subconjunto de colunas que não fazem parte da chave primária é $NK=\{VAL_SALAR, QTD_DEPEND\}$. Como ambas as colunas não possuem chave estrangeira e são de tipo numérico, os exemplos de conceitos **Measure** extraídos aplicando-se a regra R5 são:

- VAL_SALAR;
- QTD_DEPEND.

Exemplo de conceito **Measure** extraído aplicando-se a regra R6 sobre a tabela de fato FAT_ESTADO_CIVIL, mapeada do esquema da Figura 3.3 como Fato sem Fato pela regra R1, é:

- mea-FAT_ESTADO_CIVIL.

3.1.6 Classe Rollup

Um *roll up* representa a relação existente entre dois níveis de dimensão e é composto de um nível de origem e um nível de destino. Existem 2 cenários para o mapeamento de *roll ups*.

No cenário 1, um *roll up* representa a relação existente entre dois níveis de dimensões diferentes através de chave estrangeira. O nível de dimensão da tabela com chave primária referenciada é o destino no *roll up* e o outro nível de dimensão, da tabela

que contém a chave estrangeira, é a origem. Por exemplo, o *roll up* do nível de dimensão DIM_IDADE para o nível de dimensão DIM_FAIXA_ETARIA, contidos nas tabelas com os mesmos respectivos nomes na Figura 3.3, e nomeado ROLLUP_IDADE_FAIXA_ETARIA. Para este cenário, caso as dimensões envolvidas no *roll up* apresentem mais de um nível, será considerado o nível com a mesma cardinalidade da dimensão; ou seja, mesma quantidade de registros possíveis.

No cenário 2, um *roll up* representa a relação existente entre dois níveis pertencentes a uma mesma dimensão do DW. Nesse caso, o nível com menor cardinalidade, ou seja, com menor quantidade de registros distintos, é o destino no rollup; o outro nível é a origem. Por exemplo, o *roll up* do nível de dimensão MES para o nível de dimensão ANO, ambos contidos na tabela DIM_TEMPO da Figura 3.3, e nomeado ROLLUP_MES_ANO. São encontrados menos valores distintos para ANO que para MES.

Formalmente, os conceitos **RollUp** são extraídos aplicando-se as regras R7 e R8 definidas nas Figuras 3.10 e 3.11, respectivamente:

Regra R7: Para cada dimensão $D1$, seja $FK=\{C_i, \dots, C_j\}$ o subconjunto de colunas de $D1$ que contenham chaves estrangeiras, e para cada dimensão $D2$, seja $PK=\{C_m, \dots, C_n\}$ o subconjunto de colunas de $N2$ que compõem a sua chave primária. A constraint de uma coluna C_a ($C_a \in FK$) referenciando C_b ($C_b \in PK$) é mapeada para um *roll up* $R1$. O nível $N1$ de origem é o nível que pertence a $D1$ e apresenta sua mesma cardinalidade e o nível $N2$ de destino é o nível que pertence a $D2$ e apresenta sua mesma cardinalidade.

Figura 3.10: Regra R7 - Mapeamento de conceitos RollUp

Regra R8: Para cada dupla de níveis de dimensão $N1$ e $N2$ mapeados da tabela $T1$, caso $N1 \neq N2$ é criada uma instância de *roll up* $R1$. Seja $VN1=\{V_1, \dots, V_i\}$ o conjunto de valores distintos de $N1$ e $VN2=\{V_1, \dots, V_j\}$ o conjunto de valores distintos de $N2$, caso $|VN1| \geq |VN2|$, $N1$ é o nível de origem. Caso $|VN1| < |VN2|$, $N2$ é o nível de origem. O nível restante é o destino.

Figura 3.11: Regra R8 - Mapeamento de conceitos RollUp

Exemplos de conceitos **RollUp** extraídos aplicando-se a regra R7 sobre o esquema da Figura 3.3 são:

- ROLLUP_FUNCIONARIO_IDADE;
- ROLLUP_IDADE_FAIXA_ETARIA.

Exemplos de conceitos **RollUp** extraídos aplicando-se a regra R8 sobre o esquema da Figura 3.3 são:

- ROLLUP_TEMPO_MES;
- ROLLUP_MES_ANO.

3.1.7 Classe Hierarchy

Uma hierarquia é formada pelo conjunto de *roll ups* que possuam níveis de dimensão em comum. Em cada *roll up* de uma hierarquia, seu nível de origem é o nível de destino de outro *roll up* ou seu nível de destino é o nível de origem de outro *roll up*. Existem 2 cenários para o mapeamento de hierarquias, semelhantes aos cenários descritos para extração dos conceitos **RollUp**.

O cenário 1 é a identificação dos *roll ups* da hierarquia através da relação por chave estrangeira, ou seja, incluindo *roll ups* que seus níveis de dimensão se relacionem através de chave estrangeira. O menor nível nessa hierarquia é o *roll up* com nível de origem contido em uma dimensão relacionada a uma tabela de fato e o maior nível na hierarquia é o *roll up* com nível de destino contido em uma dimensão que não possua chave estrangeira. Exemplificando no modelo da Figura 3.3, uma hierarquia se inicia no *roll up* DIM_FUNCIONARIO → DIM_IDADE (com DIM_FUNCIONARIO referenciada pelo fato FAT_FUNCI), que se relaciona com o *roll up* DIM_IDADE → DIM_FAIXA_ETARIA (DIM_FAIXA_ETARIA não referencia nenhuma outra tabela), sempre por chave estrangeira. Essa hierarquia é nomeada HIERARCHY_FUNCIONARIO_IDADE_FAIXA_ETARIA, devido aos níveis de dimensão envolvidos.

O cenário 2 é a identificação dos *roll ups* com todos os níveis de uma mesma dimensão. A subordinação entre os *roll ups* é definida pela cardinalidade dos níveis: quanto menor a cardinalidade, maior sua posição na hierarquia. Também exemplificando no modelo da Figura 3.3, uma hierarquia na dimensão DIM_TEMPO se inicia no *roll up* DIM_TEMPO → MES (DIM_TEMPO com maior cardinalidade), que se relaciona com o *roll up* MES → ANO (ANO com menor cardinalidade). Essa hierarquia é nomeada HIERARCHY_TEMPO, devido ao nome da dimensão que contém os níveis.

Formalmente, os conceitos **Hierarchy** são extraídos aplicando-se as regras R9 e R10 definidas na Figura 3.12 e 3.13, respectivamente:

Regra R9: Seja $HN=\{R_i, \dots, R_j\}$ um conjunto de rollups, NO_i o nível de origem do *roll up* R_i e ND_i o nível de destino do *roll up* R_i . HN é mapeado para uma hierarquia $H1$ se, $\forall a$, NO_a e ND_a ($R_a \in HN$) se relacionam através de chave estrangeira e $\exists R_b$ ($R_b \in HN$) onde $NO_a = ND_b$ ou $NO_b = ND_a$. Seja $FN=\{F_k, \dots, F_l\}$ o conjunto de todos os fatos do DW, o menor nível em $H1$ é o rollup R_m ($R_m \in H1$), onde NO_m é referenciado por alguma coluna de F_m ($F_m \in FN$) através de chave estrangeira, e o maior nível em $H1$ é o rollup R_n ($R_n \in H1$), onde, $\forall p$, seja $R_p \in H1$, $\nexists ND_n = NO_p$.

Figura 3.12: Regra R9 - Mapeamento de conceitos **Hierarchy**

Regra R10: Seja $HN1=\{R_i, \dots, R_j\}$ o conjunto de rollups entre os níveis da dimensão $D1$. $HN1$ é mapeado para uma hierarquia $H1$ se $|HN1| > 0$. Seja $ND1=\{N1, \dots, Nn\}$ o conjunto de níveis de dimensão da dimensão $D1$, $VNa=\{V1, \dots, Vj\}$ e $VNb=\{V1, \dots, Vk\}$ o conjunto de valores distintos de Na ($Na \in ND1$) e Nb ($Nb \in ND1$), respectivamente. Caso $|VNa| > |VNb|$, Nb é superior a Na em $H1$. Caso $|VNa| \leq |VNb|$, Na é superior a Nb em $H1$.

Figura 3.13: Regra R10 - Mapeamento de conceitos **Hierarchy**

Exemplo de conceito **Hierarchy** extraído aplicando-se a regra R9 sobre o esquema da Figura 3.3, conforme explicado anteriormente, é:

- HIERARCHY_FUNCIONARIO_IDADE_FAIXA_ETARIA;

Exemplo de conceito **Hierarchy** extraído aplicando-se a regra R10 sobre o esquema da Figura 3.3, conforme explicado anteriormente, é:

- HIERARCHY_TEMPO;

3.1.8 Classe SummarizabilityAlongFact

A agregabilidade ao longo de fatos de uma medida $M1$ representa todas as relações que $M1$ tem com as tabelas de fato do esquema de dados. A estratégia é identificar todas as tabelas de fato que implementem uma medida com o mesmo nome $M1$. Como as medidas são implementadas nas tabelas de fato através de colunas, em todas as tabelas de fato, uma mesma medida $M1$ será implementada sempre com uma coluna de mesmo nome. O nome do conceito é definido com a concatenação do prefixo “saf-” e o nome da medida.

Formalmente, os conceitos **SummarizabilityAlongFact** são extraídos aplicando-se a regra R11 definida na Figura 3.14:

Regra R11: Seja $FN=\{F_1, \dots, F_j\}$ o conjunto de todos os fatos do DW. Para cada medida M_1 , é mapeada uma instância de agregabilidade ao longo de fatos AF_1 de M_1 , relacionada a M_1 e ao fato F_a ($F_a \in FN$), caso F_a contenha a medida M_1 .

Figura 3.14: Regra R11 - Mapeamento de conceitos **SummarizabilityAlongFact**

Como exemplo, a medida VAL_SALAR mapeada pela regra R5 é encontrada em mais de uma tabela de fato do esquema da Figura 3.3; a coluna VAL_SALAR que a representa é encontrada nas tabelas FAT_FUNCI, AGR_FUNCI_ANO e AGR_FUNCI_FAIXA_ETARIA. Assim, com a aplicação da regra R11 sobre o esquema da Figura 3.3, é mapeado o seguinte conceito **SummarizabilityAlongFact**, relacionado à medida VAL_SALAR e às tabelas de fato FAT_FUNCI, AGR_FUNCI_ANO e AGR_FUNCI_FAIXA_ETARIA:

- saf-VAL_SALAR.

3.1.9 Classe **SummarizabilityAlongDimension**

A agregabilidade ao longo de dimensões de uma medida M_1 representa todas as relações que M_1 tem com as dimensões do esquema de dados. Essas relações são identificadas através das tabelas de fato que contém M_1 . Toda dimensão que se relaciona com uma dessas tabela de fato, também foi definido que se relaciona com M_1 . Assim, são consultadas as tabelas de fato relacionadas aos conceitos **SummarizabilityAlongFact** já mapeados anteriormente e identificadas as dimensões relacionadas a esses fatos. O nome do conceito é definido com a concatenação do prefixo “sad-” e o nome da medida.

Formalmente, os conceitos **SummarizabilityAlongDimension** são extraídos aplicando-se a regra R12 definida na Figura 3.15:

Regra R12: Seja $F_1=\{F_1, \dots, F_j\}$ o conjunto de todas as tabelas de fato que contém a medida M_1 e $D=\{D_m, \dots, D_n\}$ o conjunto de todas as dimensões do DW. Para cada medida M_1 , é mapeada uma instância de agregabilidade ao longo de dimensões AD_1 de M_1 , relacionada a M_1 e à dimensão D_1 ($D_1 \in D$), caso D_1 se relacione com alguma tabela de fato F_a ($F_a \in F_1$) através de chave estrangeira.

Figura 3.15: Regra R12 - Mapeamento de conceitos **SummarizabilityAlongDimension**

Como exemplo, a medida VAL_SALAR é encontrada nas tabelas de fato FAT_FUNCI, AGR_FUNCI_ANO e AGR_FUNCI_FAIXA_ETARIA. Essas tabelas se relacionam a diferentes dimensões por chave estrangeira: FAT_FUNCI e AGR_FUNCI_ANO se relacionam com DIM_TEMPO e DIM_FUNCIONARIO, e AGR_FUNCI_FAIXA_ETARIA se relaciona com DIM_TEMPO e DIM_FAIXA_ETARIA. Assim, com a aplicação da regra R12 sobre o esquema da Figura 3.3, é mapeado o seguinte conceito **SummarizabilityAlongDimension**, relacionado à medida VAL_SALAR e às dimensões DIM_TEMPO, DIM_FUNCIONARIO e DIM_FAIXA_ETARIA:

- sad-VAL_SALAR.

3.1.10 Classe SummarizabilityAlongHierarchy

A agregabilidade ao longo de hierarquias de uma medida M1 representa as relações que M1 tem com hierarquias mapeadas. Essas relações são identificadas através das dimensões que se relacionam com M1. Caso uma dimensão que se relaciona M1 tenha algum nível que pertença a uma hierarquia, essa hierarquia também se relaciona com M1. Assim, são consultadas as dimensões relacionadas aos conceitos **SummarizabilityAlongDimension** já mapeados anteriormente e identificadas as hierarquias relacionadas a essas dimensões. O nome do conceito é definido com a concatenação do prefixo “sah-” e o nome da medida.

Formalmente, os conceitos **SummarizabilityAlongHierarchy** são extraídos aplicando-se a regra R13 definida na Figura 3.16:

Regra R13: Seja $D1=\{D_i, \dots, D_j\}$ o conjunto de todas as dimensões relacionadas à medida M1 e $H=\{H_i, \dots, H_j\}$ o conjunto de todas as hierarquias mapeadas do DW. Para cada medida M1, é mapeada uma instância de agregabilidade ao longo de hierarquias AH1 de M1, relacionada a M1 e à hierarquia H1 ($H1 \in H$), caso H1 tenha algum *roll up* com nível de dimensão mapeado a partir de alguma dimensão D_a ($D_a \in D1$).

Figura 3.16: Regra R13 - Mapeamento de conceitos **SummarizabilityAlongHierarchy**

Como exemplo, a medida VAL_SALAR é relacionada às dimensões DIM_TEMPO, DIM_FUNCIONARIO e DIM_FAIXA_ETARIA. Essas dimensões contêm *roll ups* que pertencem a hierarquias mapeadas: DIM_TEMPO se relaciona com

à hierarquia HIERARCHY_TEMPO, e DIM_FUNCIONARIO e DIM_FAIXA_ETARIA se relacionam com a hierarquia HIERARCHY_FUNCIONARIO_IDADE_FAIXA_ETARIA. Assim, com a aplicação da regra R13 sobre o esquema da Figura 3.3, é mapeado o seguinte conceito **SummarizabilityAlongHierarchy**, relacionado à medida VAL_SALAR e às hierarquias HIERARCHY_TEMPO e HIERARCHY_FUNCIONARIO_IDADE_FAIXA_ETARIA:

- sah-VAL_SALAR.

3.2 Transformação para OWL

As regras da OntoDW para extração de conceitos não contêm regras definidas por Prat et al. [Prat, Megdiche and Akoka, 2012]. Entretanto, para a geração da ontologia OWL após a identificação dos conceitos, algumas destas regras foram utilizadas, com ajustes. As regras aproveitadas de Prat et al. [Prat, Megdiche and Akoka, 2012] definem os conceitos identificados como subclasses das respectivas classes que representam do metamodelo de tarefa OLAP. Por exemplo, considere a Transformação T2.1 definida em [Prat, Megdiche and Akoka, 2012] e descrita na figura 3.17:

“Transformation T2.1: Each dimension of the multidimensional model is defined as a subclass of the class Dimension in the OWL-DL ontology”

Figura 3.17: Exemplo de transformação para classe na ontologia OWL [Prat, Megdiche and Akoka, 2012]

A transformação T2.1 foi utilizada no presente trabalho, com o ajuste de que cada tabela de dimensão identificada no DW foi definida como uma instância da classe **Dimension**, e não como uma subclasse. A definição dos conceitos como instâncias no arquivo OWL foi feita para melhor manipulação da ontologia, com a separação clara dos relacionamentos do modelo (instâncias) dos relacionamentos do metamodelo (classes), e pelo não uso das instâncias para representar os dados da aplicação, como os registros da tabela de dimensão, conforme previsto por Prat et al. [Prat, Megdiche and Akoka, 2012].

Apesar de todos os conceitos extraídos pelas regras de mapeamento definidos serem importantes, nem todos serão gerados na ontologia OWL. O motivo é que, para 2

casos, o metamodelo de tarefa OLAP utilizado não prevê classes ou relacionamentos para representar o conhecimento gerado.

O primeiro caso é a qualificação de um conceito **Dimension** como Dimensão de Tempo, Dimensão Versionada ou Dimensão de Negócio (regra R2). O segundo caso é definição da relação de superioridade entre *roll ups* dentro de uma hierarquia (regras R9 e R10). Em ambos os casos as regras possibilitam a geração do conhecimento, porém eles não serão explicitados na ontologia final gerada pelo OntoDW.

3.3 Implementação da solução

Para início da implementação do OntoDW foi montado um ambiente tecnológico de desenvolvimento e testes. O SGBD escolhido para o *Data Warehouse* de desenvolvimento foi o Oracle 11g Express Edition (Oracle 11g XE). Esse SGBD foi escolhido por disponibilizar uma versão gratuita sem restrições de funcionalidades que inviabilizem o estudo de caso, por ser do mesmo fabricante e versão do SGBD utilizado no *Data Warehouse* objeto do estudo de caso (apesar deste utilizar uma versão paga do SGBD) e pela experiência do autor no uso desta plataforma. Adicionalmente, o SGBD Oracle tem seu uso difundido em corporações de diferentes domínios e tamanhos em todo o mundo. O SGBD foi instalado em um Ultrabook Dell, com processador Intel i5 e 8Gb de memória RAM, sobre o sistema operacional Windows 10.

Para a realização de testes da solução proposta, era necessário que o DW de teste estivesse carregado com dados consistentes. Para essa função, foi obtida uma massa de dados de amostra com informações de funcionários e ex-funcionários num determinado período de tempo, compondo cerca de 650.000 registros. Essas informações eram a matrícula, o nome, a idade, o salário, o estado civil e a data de início do estado civil. A matrícula e o nome dos funcionários foram fornecidos com caracteres embaralhados, para preservar a identidade das pessoas. A instituição que forneceu os dados foi a mesma que seria futuramente objeto do estudo de caso.

O esquema do DW de teste foi implementado segundo o modelo ilustrado na Figura 3.3, por implementar técnicas diversas de modelagem multidimensional e ser aderente à massa de dados de teste. A carga dos dados foi realizada após a execução das etapas de limpeza e tratamento (ETL), implementadas em scripts PL/SQL (linguagem

procedural proprietária da Oracle). As tabelas de dimensão foram carregadas primeiramente, gerando desde tabelas com baixa cardinalidade (como a dimensão Estado Civil) a tabelas com cardinalidade muito alta (como a dimensão Funcionário). Em seguida, foi realizada a carga das tabelas de fato. Os scripts de criação e de carga das tabelas, assim como dados da massa de testes já tratados para carga, estão disponíveis em <https://sourceforge.net/projects/ontodw/files/implementacao>.

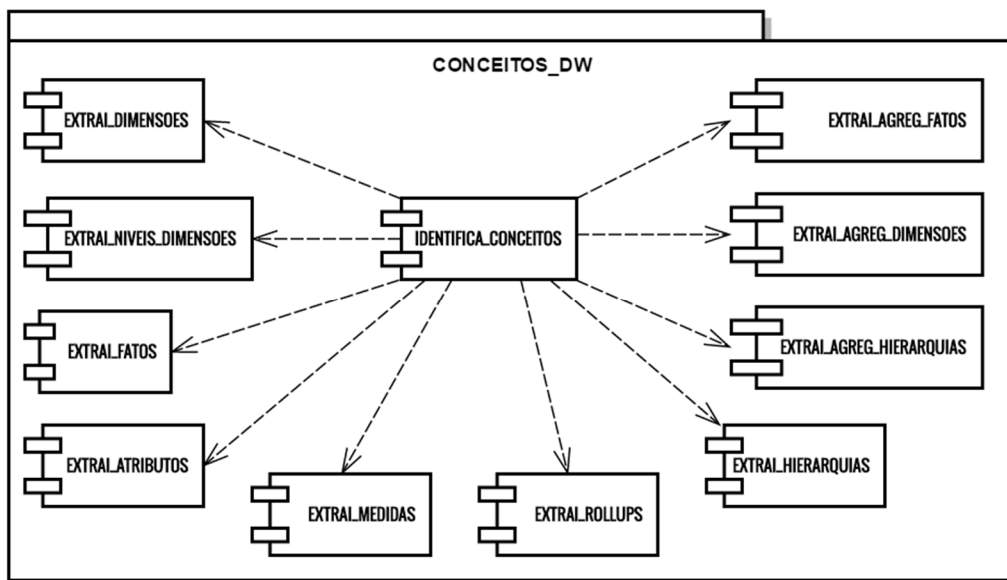


Figura 3.18: Diagrama de componentes da implementação do OntoDW

Como metamodelo de domínio para o ambiente de desenvolvimento, foi definido que, para os nomes das estruturas do esquema do DW, cada termo é separado por um símbolo “_”. Para as colunas, o primeiro termo encontrado é definidor do tipo de conteúdo armazenado e, para as tabelas, o primeiro termo descreve a função da tabela no modelo.

Para o metamodelo de domínio, também foi criada uma tabela de glossário no banco de dados para armazenamento de termos, com tamanho máximo de 6 caracteres, e seus respectivos significados. Dentre os termos carregados, estão palavras pertencentes ao domínio da aplicação (por exemplo, SALAR, significando Salário) e qualificadores dos objetos do banco de dados (por exemplo DIM e VAL, significando Dimensão e Valor, respectivamente). Foi também desenvolvida uma função para leitura do metamodelo de domínio, de forma a facilitar a adaptação a outros esquemas e metamodelos de domínio, bastando ajustar esta função ao ambiente de execução.

As regras de mapeamento apresentadas na seção 3.1 foram implementadas em programas PL/SQL e encapsuladas em uma *package* (chamada CONCEITOS_DW) para melhor organização e portabilidade, conforme ilustrado na Figura 3.18. Para cada classe a ser mapeada foi desenvolvido um programa para geração de suas instâncias. Desta forma, não seria necessária a preparação do banco de dados em caso de execução da solução em diferentes organizações ou utilizando diversos esquemas de um mesmo DW (apenas parametrizações, como o nome dos esquemas a serem lidos e caracteres separadores) ou a instalação de softwares desenvolvidos, que poderia incorrer em dificuldades, como violação de políticas de segurança da informação, ou problemas, como incompatibilidade como um diferente sistema operacional, por exemplo.

Foram também definidos tipos de dados para cada conceito a ser identificado e variáveis públicas para que os programas pudessem compartilhar de forma mais fácil as instâncias descobertas por cada um deles. Estes tipos de dados definidos representam as classes do metamodelo OLAP com regras definidas para a solução, estando dessa forma também implementados também em PL/SQL.

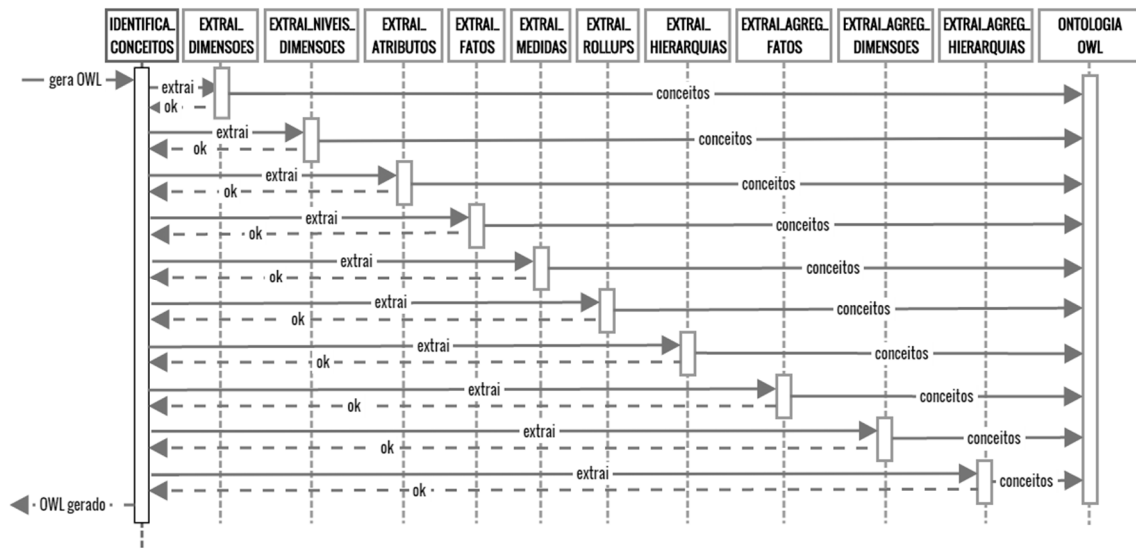


Figura 3.19: Diagrama de sequência da execução do OntoDW

No total, foram desenvolvidos 10 programas para extração dos conceitos. Nesses programas foram implementadas as regras de mapeamento definidas e utilizados os dados e metadados lidos do DW, as variáveis compartilhadas e o glossário como fontes de informação. O resultado da execução sequencial desses programas é a geração de um arquivo OWL contendo a ontologia de aplicação esperada. A execução dos programas na sequência correta foi implementada em um programa na mesma *package* (chamado

IDENTIFICA_CONCEITOS), de forma a facilitar a geração da ontologia e evitar erros na sequência de execução, conforme ilustrado na Figura 3.19. A versão final da *package* Oracle desenvolvida está disponível em <https://sourceforge.net/projects/ontodw/files/estudo%20de%20caso/package%20Oracle>.

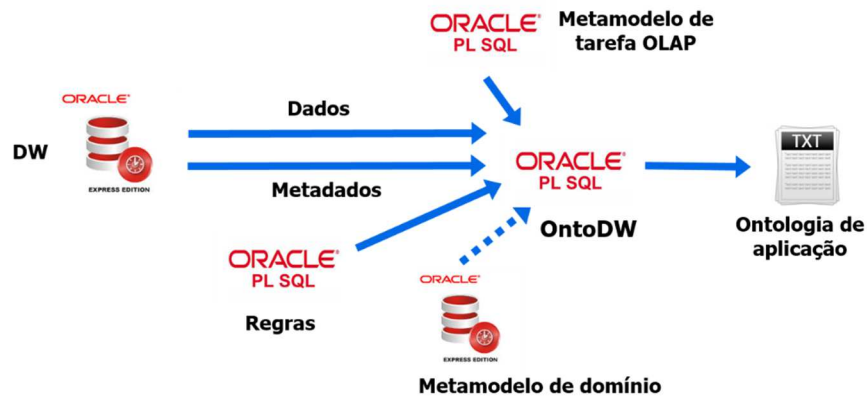


Figura 3.20: Arquitetura do ambiente de desenvolvimento

A visão geral da arquitetura do ambiente de criado para desenvolvimento da solução é ilustrada na Figura 3.20, onde cada componente é apresentado com a tecnologia utilizada.

Tabela 3.2: Níveis de dimensão mapeados do DW de teste

| Nível | Valores distintos | Dimensão |
|------------------|-------------------|------------------|
| DIM_FAIXA_ETARIA | 9 | DIM_FAIXA_ETARIA |
| DIM_FUNCIONARIO | 275.267 | DIM_FUNCIONARIO |
| DIM_ESTADO_CIVIL | 8 | DIM_ESTADO_CIVIL |
| DIM_TEMPO | 5.481 | DIM_TEMPO |
| MES | 182 | DIM_TEMPO |
| ANO | 17 | DIM_TEMPO |
| DIM_IDADE | 104 | DIM_IDADE |

A implementação da regra para extração dos níveis de dimensão (Regra R3) foi a que representou maior complexidade. Ela utiliza os dados armazenados nas tabelas de dimensão, além dos metadados, e compara os dados nas colunas das tabelas. A estratégia de implementação adotada para esta regra foi a comparação de todas as colunas em pares, buscando a existência de mais de um valor possível na coluna B para cada valor da coluna A e vice-versa. Caso não fosse encontrada nenhuma ocorrência em nenhuma das duas consultas podíamos afirmar que pertenciam a um mesmo nível de dimensão. Se uma das colunas já pertencia a um nível identificado, a outra também era incluída no

mesmo nível. Essa estratégia tornou o programa com menor custo de processamento e com código-fonte menor e mais claro.

Tabela 3.3: Agregabilidades de medidas com fatos mapeadas do DW de teste

| Medida | Fato |
|------------------------|------------------------------|
| VAL_SALAR | AGR_FUNCIONARIO |
| | AGR_FUNCIONARIO_FAIXA_ETARIA |
| | FAT_FUNCIONARIO |
| QTD_DEPEND | AGR_FUNCIONARIO |
| | AGR_FUNCIONARIO_FAIXA_ETARIA |
| | FAT_FUNCIONARIO |
| COUNT_FAT_ESTADO_CIVIL | FAT_ESTADO_CIVIL |

A análise dos resultados foi realizada para cada regra (tipos de conceitos mapeados), considerando uma regra bem-sucedida caso ela tenha identificado na ontologia gerada todas as instâncias esperadas.

Tabela 3.4: Agregabilidades de medidas com dimensões mapeadas do DW de teste

| Medida | Dimensão |
|------------------------|------------------|
| VAL_SALAR | DIM_TEMPO |
| | DIM_FAIXA_ETARIA |
| | DIM_FUNCIONARIO |
| QTD_DEPEND | DIM_TEMPO |
| | DIM_FAIXA_ETARIA |
| | DIM_FUNCIONARIO |
| COUNT_FAT_ESTADO_CIVIL | DIM_ESTADO_CIVIL |
| | DIM_TEMPO |

O primeiro teste realizado não foi bem-sucedido na extração dos níveis da dimensão DIM_TEMPO, única tabela desnormalizada do esquema do DW de avaliação. Após análise preliminar, foi identificado que o motivo do não funcionamento foi o uso de códigos reservados na dimensão para representar registros na tabela fato onde aquela dimensão não seria pertinente, prática comum em implementação de *Data Warehouses* para não utilização de valores nulos. No DW de teste foram utilizados 2 registros com códigos reservados, um para representar os casos onde um valor para a dimensão não se aplica e outro para representar os casos onde o valor não foi informado ou encontrado na carga. No entanto, os códigos reservados foram utilizados em uma das colunas também para outros registros. O programa identificou um nível de dimensão a mais para esta coluna.

Tabela 3.5: Agregabilidades de medidas com hierarquias mapeadas do DW de teste

| Medida | Hierarquia |
|------------|-------------------|
| VAL_SALAR | DIM_TEMPO_DESNORM |
| | DIM_FUNCIONARIO |
| QTD_DEPEND | DIM_TEMPO_DESNORM |
| | DIM_FUNCIONARIO |

Para resolver esta falha, a implementação da regra foi ajustada para incluir a quantidade e os códigos de dimensão reservados na análise dos pares de colunas. Foi criada uma constante global para parametrização desta quantidade, que é utilizada como uma espécie de margem de erro, e parâmetros de códigos reservados. Caso as colunas sejam parte da chave primária da tabela, os valores dos códigos reservados são também incluídos na análise, de forma a desconsiderar os registros reservados na comparação com os dados das colunas.

Foi realizada nova execução dos programas e extraídos novos grupos de conceitos. Desta vez foram obtidos os níveis de dimensão esperados, conforme apresentado na Tabela 3.2. Para avaliação do resultado, os conceitos gerados foram comparados aos conceitos identificados de forma visual pelo autor, especialista da área de *Business Intelligence*, utilizando somente o modelo físico do DW ilustrado na da Figura 3.3. Para as dimensões com apenas 1 nível, o nome do nível é o mesmo da dimensão. Quando a dimensão apresenta mais de um nível, o nível de menor granularidade tem o mesmo nome da dimensão. Os níveis restantes têm seus nomes definidos através de consulta ao glossário com os termos obtidos a partir do nome do atributo definido como identificador do nível.

A Tabela 3.3 apresenta a agregabilidade das medidas identificadas no DW com as tabelas fato onde elas estão inseridas. A Tabela 3.4 apresenta a agregabilidade das medidas com as dimensões que se relacionam com as tabelas fato onde elas estão inseridas. A Tabela 3.5 apresenta a agregabilidade das medidas com as hierarquias identificadas. Quanto às hierarquias, elas são identificadas como agregáveis com uma medida se qualquer *roll up* que a compõe tiver dimensão relacionada com uma tabela fato que tenha a medida.

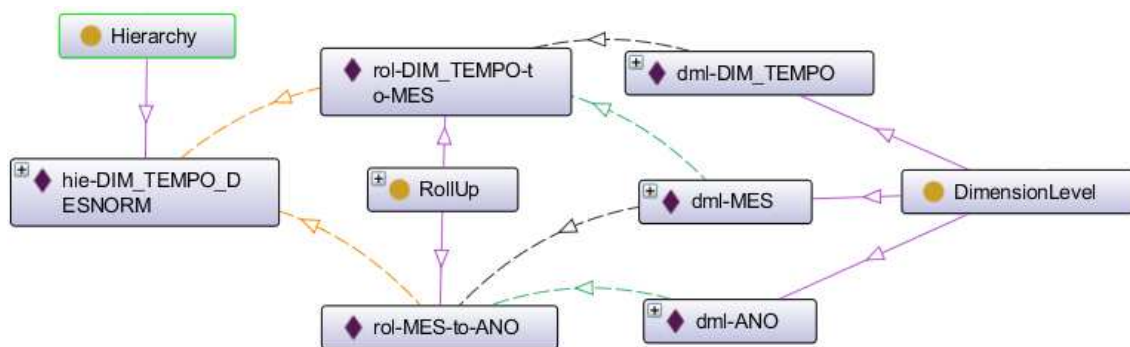


Figura 3.21: Recorte da ontologia extraída do ambiente de testes

Após a finalização do desenvolvimento dos programas, foi realizada uma execução completa dos programas do OntoDW no ambiente de testes e extraída sua ontologia correspondente. Para visualização da ontologia gerada foi utilizado o software Protégé (www.protege.stanford.edu/). O Protégé é um editor e visualizador gratuito e open source de ontologias, que provê uma interface gráfica aos usuários para manipulá-las. Na Figura 3.21 é apresentado um recorte desta ontologia, visualizada através do plugin OntoGraf do Protégé. Neste recorte podemos observar uma hierarquia identificada (hie-DIM_TEMPO_DESNORM), os dois *roll ups* que a compõem (rol-DIM_TEMPO-to-MES e rol-MES-to-ANO) e os níveis de dimensão que compõem cada *roll up* (dml-DIM_TEMPO, dml-MES e dml-ANO). Cada *roll up* tem um nível de origem e um nível de destino.

Na Figura 3.21 foram mantidas também as classes correspondentes às instâncias, para melhor entendimento do modelo. Todas as instâncias apresentadas no recorte foram extraídas a partir da tabela DIM_TEMPO e exemplificam o conhecimento implícito que existe no *Data Warehouse*.

O resultado obtido com a execução do OntoDW no ambiente de testes foi satisfatório, por identificar conceitos pertinentes ao modelo implementado no DW, mesmo que não sendo o resultado 100% conforme o esperado. Com a descoberta dos conceitos conforme esperado, o próximo passo foi a execução do estudo de caso previsto.

4 – Estudo de Caso

Este capítulo apresenta o estudo de caso executado para avaliação da proposta, incluindo o cenário de aplicação e detalhes do projeto de avaliação.

4.1 Cenário de aplicação

Para avaliação da proposta de solução foi conduzido um estudo de caso, no cenário de um fundo de pensão dos funcionários de uma das maiores instituições financeiras da América Latina. Esta instituição financeira é centenária e tem em seu quadro atualmente mais de 100.000 funcionários e oferece a seus clientes mais de 15.000 pontos de atendimento espalhados pelo Brasil. O fundo de pensão em questão também é uma das maiores instituições da América Latina, com a responsabilidade de administrar recursos financeiros na ordem de dezenas de bilhões de reais e pagar mensalmente centenas de milhões de reais em benefícios.

O domínio escolhido compreende o cadastro de funcionários, escolhido neste trabalho por ser um domínio de mais fácil entendimento para o público não especializado, pois a maioria dos conceitos são de conhecimento comum a funcionários de qualquer organização; além disso, o esquema do DW correspondente a este domínio na organização aplica uma diversidade de técnicas modelagem multidimensional e os dados armazenados no DW são conhecidos por sua consistência; por fim, ele é um assunto estratégico para a área de *Business Intelligence* da instituição e para a área de negócio gestora de seus dados.

Este banco de dados tem a finalidade de disponibilizar informações aos usuários para apoio à tomada de decisão e execução dos processos de negócio, seja através da ferramenta OLAP existente ou de relatórios agendados para a geração de planilhas. O público-alvo é diverso, atendendo desde a analistas das áreas de negócio quanto ao presidente e diretores da organização, além do público externo. Os dados são acessados de forma on-line ou por cache gerado, seja através da rede corporativa existente via browser, no site da organização ou através dispositivos móveis (*iPads* e *iPhones*) a

qualquer momento. Algumas dessas informações são públicas e outras com acesso restrito. Por estes motivos, esse ambiente dispõe de alta disponibilidade. Ao longo dos anos, essa aplicação sofreu manutenções corretivas e evolutivas para se manter alinhada às necessidades do negócio sem a correspondente atualização da documentação existente.

O DW armazena dados carregados desde 1997, totalizando dezenas de milhões de registros nas tabelas. Esses dados são relativos a 275.000 funcionários e ex-funcionários da instituição financeira, sob diversas visões de análise, compondo um ambiente rico de informações que pode ser utilizado para ações gerenciais e análises relativas a cálculo atuarial e acompanhamento do corpo funcional. Mensalmente, as informações cadastrais dos funcionários disponibilizadas pela instituição financeira são carregadas neste DW e integradas com informações do fundo de pensão a respeito do vínculo desses funcionários com os planos de previdência o qual podem ser participantes.

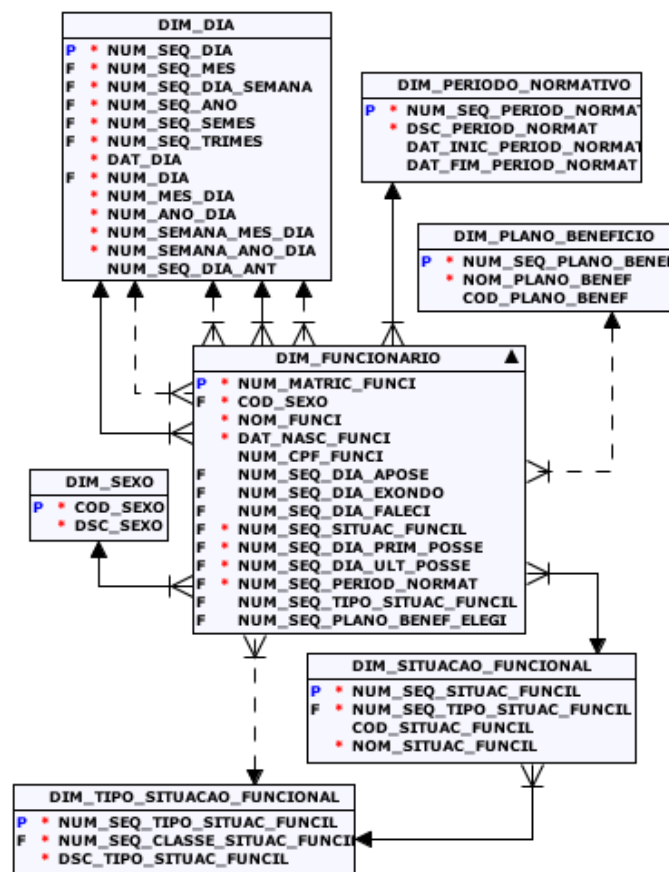


Figura 4.1: Recorte do modelo do esquema do DW do estudo de caso com dimensões relacionadas ao conceito Funcionário

Este esquema do DW contém 62 tabelas (50 tabelas de dimensão, 11 tabelas de fato e uma tabela de controle), e a mais populosa delas armazena aproximadamente 55 milhões de registros. O DW está implementado sobre o SGBD Oracle 11g.

Na Figura 4.1 é apresentado um recorte do modelo físico de dados deste esquema. É apresentada a tabela de dimensão DIM_FUNCIONARIO e algumas outras dimensões que se relacionam com ela através de chave estrangeira. Essas outras dimensões são potenciais agrupamentos de funcionários que podem ser utilizados em uma análise.

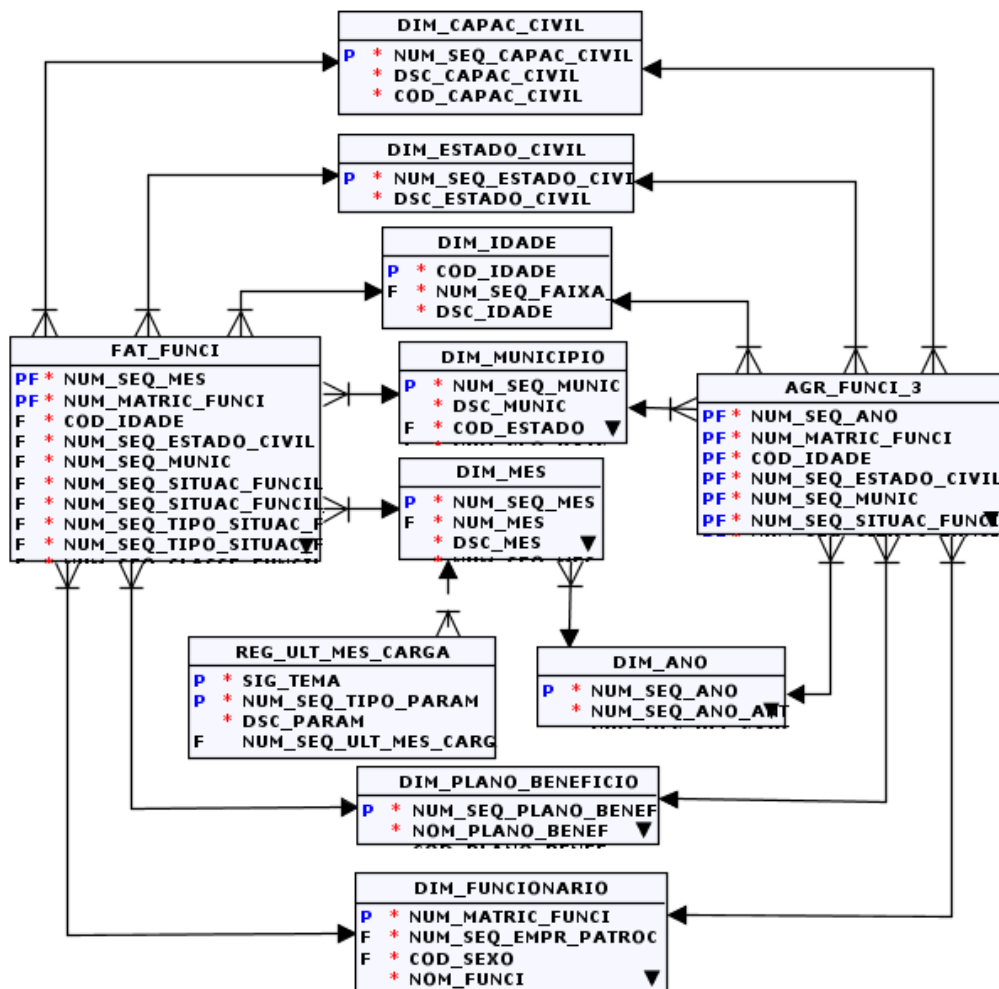


Figura 4.2: Recorte do modelo do esquema do DW do estudo de caso com fatos e dimensões

Na Figura 4.2 é apresentado outro recorte do mesmo modelo físico de dados. Existem 2 tabelas de fato/agregação (FAT_FUNCNI e AGR_FUNCNI_3) e algumas das tabelas de dimensão existentes que se relacionam com elas, armazenando visões de análise.

lise das medidas/métricas disponíveis. Uma tabela de controle também está presente (REG_ULT_MES_CARGA), que mantém o rastreamento de quais dados já estão carregados neste esquema do DW.

| AGR_FUNC1_3 | FAT_FUNC1 |
|------------------------------------|-----------------------------------|
| PF * NUM_SEQ_ANO | PF * NUM_SEQ_MES |
| PF * NUM_MATRIC_FUNC1 | PF * NUM_MATRIC_FUNC1 |
| PF * COD_IDADE | F * COD_IDADE |
| PF * NUM_SEQ_ESTADO_CIVIL | F * NUM_SEQ_ESTADO_CIVIL |
| PF * NUM_SEQ_MUNIC | F * NUM_SEQ_MUNIC |
| PF * NUM_SEQ_SITUAC_FUNC1 | F * NUM_SEQ_SITUAC_FUNC1 |
| PF * NUM_SEQ_SITUAC_FUNC1_ANT | F * NUM_SEQ_SITUAC_FUNC1_ANT |
| PF * NUM_SEQ_TIPO_SITUAC_FUNC1 | F * NUM_SEQ_TIPO_SITUAC_FUNC1 |
| PF * NUM_SEQ_TIPO_SITUAC_FUNC1_ANT | F * NUM_SEQ_TIPO_SITUAC_FUNC1_ANT |
| PF * NUM_SEQ_CLASSE_FUNC1_ANT | F * NUM_SEQ_CLASSE_FUNC1_ANT |
| PF * NUM_SEQ_DEPEND | F * NUM_SEQ_DEPEND |
| PF * NUM_SEQ_COMIS | F * NUM_SEQ_COMIS |
| PF * NUM_SEQ_VP | F * NUM_SEQ_VP |
| PF * COD_AN | F * COD_AN |
| PF * NUM_SEQ_PARTIC | F * NUM_SEQ_PARTIC |
| PF * NUM_SEQ_PLANO_BENEF | F * NUM_SEQ_PLANO_BENEF |
| PF * NUM_SEQ_SITUAC_PREVI | F * NUM_SEQ_SITUAC_PREVI |
| PF * NUM_SEQ_STATUS_PREVI | F * NUM_SEQ_STATUS_PREVI |
| PF * NUM_SEQ_PERIOD_NORMAT | F * NUM_SEQ_TARIFA |
| PF * NUM_SEQ_EMPR_PATROC | F * NUM_SEQ_FAIXA |
| PF * NUM_SEQ_CAPAC_CIVIL | F * NUM_SEQ_NIVEL |
| PF * NUM_SEQ_TIPO_RENDA_PROGDA | F * NUM_SEQ_SUB_GRUPO |
| PF * NUM_SEQ_STATUS_BB | F * NUM_SEQ_UNID_ORGANZ_BB |
| P * IND_RECIBTO_APOSE | F * NUM_SEQ_MEDIDA |
| * QTD_FUNC1 | F * NUM_SEQ_MEDIDA_MOTIVO |
| * QTD_DEP | F * NUM_SEQ_PERIOD_NORMAT |
| * QTD_BENEF | F * NUM_SEQ_EMPR_PATROC |
| * QTD_PENSTA | F * NUM_SEQ_CAPAC_CIVIL |
| * QTD_FILIA | F * NUM_SEQ_TIPO_RENDA_PROGDA |
| * QTD_DESFIL | * QTD_FUNC1 |
| * QTD_REINGR | * QTD_DEP |
| * VAL_SALAR_PARTIC | * QTD_BENEF |
| * QTD_FUNC1_SITUAC_FUNC1_PERIOD | * QTD_PENSTA |
| * QTD_FUNC1_TIPO_FUNC1_PERIOD | * QTD_FILIA |
| * QTD_FUNC1_CLASSE_PERIOD | * QTD_DESFIL |
| * QTD_FUNC1_SITUAC_PREVI_PERIOD | * QTD_REINGR |
| * QTD_FUNC1_STATUS_PREVI_PERIOD | * VAL_SALAR_PARTIC |
| * QTD_FUNC1_COMIS_PERIOD | * QTD_FUNC1_SITUAC_FUNC1_PERIOD |
| * QTD_FUNC1_COD_VP_PERIOD | * QTD_FUNC1_TIPO_FUNC1_PERIOD |
| * QTD_FUNC1_AN_PERIOD | * QTD_FUNC1_CLASSE_PERIOD |
| * QTD_FUNC1_PLANO_BENEF_PERIOD | * QTD_FUNC1_SITUAC_PREVI_PERIOD |
| * NUM_DIA_CONTRI_INSS_FORA | * QTD_FUNC1_STATUS_PREVI_PERIOD |
| * NUM_DIA_CONTRI_INSS_PATROC | * QTD_FUNC1_COMIS_PERIOD |
| | * QTD_FUNC1_COD_VP_PERIOD |
| | * QTD_FUNC1_AN_PERIOD |
| | * QTD_FUNC1_PLANO_BENEF_PERIOD |
| | F * NUM_SEQ_STATUS_BB |
| | * NUM_DIA_CONTRI_INSS_FORA |
| | * NUM_DIA_CONTRI_INSS_PATROC |
| | * IND_RECIBTO_APOSE |

Figura 4.3: Representação das tabelas FAT_FUNC1 e AGR_FUNC1_3

As tabelas de fato/agregação da Figura 4.2 apresentam somente um subconjunto de suas colunas, pois o número total de colunas nas tabelas é muito alto (50 colunas para a tabela FAT_FUNC1). Isto se deve ao número alto de dimensões e métricas existentes para este assunto no DW e também pelo fato de que se trata de uma estrutura de dados antiga, que já sofreu diversas alterações/manutenções evolutivas e corretivas. Na Figura 4.3 é possível observar a representação completa das tabelas FAT_FUNC1 e AGR_FUNC1_3, para exemplificar esta questão, que torna mais difícil para analistas a realização de análises do modelo para identificação das informações existentes no DW. Por exemplo, a medida Salário é armazenada pela coluna VAL_SALAR_PARTIC. Entre-

tanto, ela está presente nestas duas tabelas com uma identificação confusa e diferentes visões de análise possíveis a partir do modelo de dados.

4.2 Projeto de avaliação

Para a avaliação do estudo de caso, o plano foi avaliar cada regra individualmente para confirmar sua capacidade de identificação dos conceitos previstos. Uma tentativa de avaliação individual da ontologia gerada poderia depender excessivamente de conhecimento do domínio pelo avaliador e seria difícil identificar as causas de possíveis problemas encontrados. Outra questão é que uma avaliação apenas pelo autor poderia ocasionar em um resultado influenciado pelo conhecimento que o mesmo possui na estrutura de dados do estudo de caso e nas regras de mapeamento definidas. Assim sendo, a decisão foi de realizar a avaliação individual das regras de mapeamento com especialistas da área de *Business Intelligence*, com conhecimento e experiência em modelagem multidimensional e aplicações OLAP, e de avaliar a percepção de usuários em relação à solução. A avaliação inclui 3 pesquisas com profissionais:

- I. Uma pesquisa com especialistas do mercado brasileiro de *Business Intelligence*, para comparação das respostas com as obtidas no estudo de caso;
- II. Uma pesquisa com profissionais da organização do estudo de caso, também especialistas em *Business Intelligence*, para que suas respostas sejam comparadas com as respostas obtidas no estudo de caso e do outro grupo de pesquisa;
- III. Uma pesquisa com usuários da aplicação de BI do estudo de caso, para análise de suas impressões sobre as características e a utilidade do modelo conceitual gerado.

O objetivo principal do OntoDW é a extração de conceitos de BI a partir de *Data Warehouses*. Dessa forma, a estratégia para a realização da avaliação para as pesquisas I e II será a de comparar os conceitos extraídos pela execução do OntoDW com os conceitos identificados pelos especialistas. Devido ao tamanho do esquema de dados utilizado para o estudo de caso, com um grande número de tabelas e colunas, não seria viável a identificação dos conceitos pelos participantes da avaliação para todo o modelo, assim como a comparação desses resultados. Para viabilizar a realização desta avalia-

ção, a solução foi elencar trechos do esquema de dados de onde pudessem ser extraídos conceitos de todas as classes definidas. Dessa forma, foram extraídos recortes do esquema de dados do DW e criadas questões com os mesmos para registro dos conceitos identificados pelos respondentes.

Os formulários criados para as pesquisas I e II apresentam pelo menos uma questão para cada regra definida e contém recortes do modelo de dados do *Data Warehouse* utilizado no estudo de caso e informações adicionais para auxílio no entendimento. O participante da pesquisa analisará a questão e informará os conceitos que identificar utilizando os trechos de modelo fornecidos e seu conhecimento teórico e sua experiência relativos a modelagem multidimensional e aplicações de *Business Intelligence*. Nos questionário também estão definidos os conceitos que estão sendo tratados na pesquisa, para explicitar o que está sendo solicitado nas questões e mitigar o risco de existir dúvidas dos respondentes.

A estimativa inicial do número de participantes para o grupo de especialistas do mercado foi de 20 pessoas. A expectativa para o número de participantes não foi maior devido ao fato de que a área de *Business Intelligence* não é uma das mais abrangida entre os profissionais no mercado brasileiro de TI, sendo uma espécie de nicho, e pelo grau de dificuldade existente para o preenchimento da pesquisa, que não é trivial por envolver modelagem conceitual, conceitos de aplicações OLAP e análise de modelos multidimensionais físicos de dados.

Para análise dos resultados, será considerada como premissa que as respostas obtidas pelas pesquisas devem ser as mesmas fornecidas pelo OntoDW, visto que a tarefa executada é a mesma. Partindo deste ponto, serão realizadas as análises sobre as possíveis divergências encontradas e outras observações sobre os resultados.

O grupo de profissionais de TI da organização respondente da pesquisa II também dever ter conhecimento e experiência em modelagem multidimensional e aplicações OLAP para estar apto a responder a pesquisa. A comparação entre as respostas dos dois grupos de profissionais e os comentários colhidos podem trazer informações adicionais para a análise dos resultados. A quantidade prevista de participantes para esta pesquisa é de 6 pessoas. Esse número é condizente com a quantidade de profissionais de TI na organização aptos a responder a pesquisa.

Os diferenciais deste grupo de respondentes em relação ao grupo da pesquisa I são:

- Os profissionais terão o ambiente tecnológico da organização à disposição, além dos recortes e informações disponibilizados no formulário de pesquisa, caso julguem necessário, para acessar o esquema e ter acesso aos dados e metadados das tabelas. Dessa forma, terão à disposição os mesmos insumos utilizados pelo OntoDW para a geração da ontologia;
- O fato de trabalharem na organização do estudo de caso dá aos profissionais conhecimento do domínio, dos termos de negócio e dos padrões de nomenclatura utilizados na implementação dos sistemas de informação. Esse conhecimento pode auxiliar no entendimento dos recortes de modelos apresentados;
- As questões de pesquisa serão as mesmas apresentadas ao primeiro grupo, mas serão incluídas também questões subjetivas para que sejam colhidos os comentários e as impressões sobre as dificuldades encontradas no preenchimento do formulário ou sobre qualquer outro ponto que julgarem importante.

Por fim, a pesquisa III com usuários da aplicação de *Business Intelligence* do estudo de caso organização objetiva validar a premissa de que uma representação do conhecimento pode apoiar a análise dos dados no DW. O formulário de pesquisa apresenta questões contendo trechos com elementos da ontologia gerada no estudo de caso e informações adicionais suficientes para o entendimento das questões. O participante da pesquisa analisa as questões e informa sua opinião sobre a utilidade das representações apresentadas para o apoio das atividades que realiza com o sistema que tem à disposição. Para contextualização da pesquisa e esclarecimento de dúvidas sobre os recortes e informações apresentadas, foram realizadas previamente conversas com os usuários para informá-los do objetivo da pesquisa, explicar o tipo de informação que se espera com as questões e descrever o tipo de modelo que será apresentado a eles, que pode não ser familiar por conta da notação utilizada ou pela forma como as informações foram organizadas.

O número previsto de respondentes para esta última pesquisa é de 3 profissionais, que estão envolvidos diretamente com processos críticos da organização e utilizam os dados fornecidos pelos sistemas de BI. Apesar da quantidade de usuários deste sistema na organização ser maior, o grupo de pessoas com permissão para a realização de

análises e criação de relatórios *ad hoc* na ferramenta OLAP disponível é mais restrito por questões de segurança e necessidade pelas atividades que executam. Ainda, como este formulário será totalmente composto de questões subjetivas, um número maior de respondentes poderia inviabilizar uma análise das respostas.

4.3 Execução do estudo de caso

Para início da execução do estudo de caso, a *package* desenvolvida com a implementação do OntoDW foi criada no *Data Warehouse* objeto da avaliação. Antes da execução dos programas, foi necessária a parametrização de algumas constantes da *package*:

- Separadores de termos: É possível parametrizar o caractere separador dos termos no nome das colunas e o separador utilizado para concatenar os termos ao definir um nome para uma instância de um conceito a partir do glossário. Foram definidos os caracteres “_” e “-” para esses parâmetros, respectivamente;
- Nomes e diretório dos arquivos de saída: Parâmetros do nome do arquivo com extensão OWL contendo a ontologia e do arquivo de log da execução dos programas, além da definição do diretório de saída;
- Códigos reservados de dimensão: Definição da quantidade e dos códigos reservados utilizados nas dimensões para uso na identificação dos níveis, caso existam;
- Esquemas do banco de dados: É necessário parametrizar no OntoDW os esquemas do banco de dados que contém as tabelas a serem consideradas. Esse parâmetro é importante pois o usuário que executará os programas pode ter acesso a objetos de diversos esquemas diferentes no banco de dados. Caso nulo, será considerado apenas o esquema onde a *package* está criada.

Sobre a parametrização dos nomes dos esquemas, existe uma característica importante a ressaltar. Na arquitetura do *Data Warehouse* analisado, as tabelas de dimensão dos diversos assuntos presentes são implementadas num só esquema. A justificativa é que, dessa forma, as dimensões representam conceitos únicos do negócio e são compartilhadas para análise de assuntos diversos. As tabelas de fato são divididas pelos diversos esquemas no DW, de acordo com seu assunto (*data mart*). O OntoDW foi imple-

mentado para aceitar mais de um esquema do banco de dados como parâmetro para contemplar essa possível arquitetura.

Tabela 4.1: Quantidade de instâncias de classe mapeadas no estudo de caso

| Classe | Quantidade de instâncias |
|-------------------------------|--------------------------|
| Dimension | 50 |
| DimensionLevel | 75 |
| Attribute | 144 |
| Fact | 12 |
| Measure | 103 |
| RollUp | 65 |
| Hierarchy | 23 |
| SummarizabilityAlongFact | 40 |
| SummarizabilityAlongDimension | 40 |
| SummarizabilityAlongHierarchy | 40 |
| Total | 592 |

Após a parametrização necessária, os programas estavam aptos a serem executados. Como descrito anteriormente, cada programa desenvolvido tem como objetivo identificar as instâncias de uma determinada classe e, devido à relação existente entre as classes definida pelo metamodelo de tarefa OLAP, existe uma ordem de execução dos programas que deve ser obedecida. Os programas foram executados conforme planejado e os arquivos de log e OWL foram gerados conforme esperado.

Tabela 4.2: Quantidade de relações entre instâncias mapeadas no estudo de caso

| Relação | Quantidade | Relação inversa | Quantidade |
|---------------------------|--------------|---------------------------|--------------|
| factToDimension | 163 | dimensionToFact | 163 |
| measureToFact | 103 | factToMeasure | 103 |
| dimensionToHierarchy | 226 | hierarchyToDimension | 226 |
| rollUpToHierarchy | 113 | hierarchyToRollUp | 113 |
| isSourceOf | 65 | hasSource | 65 |
| isTargetOf | 65 | hasTarget | 65 |
| attributeToDimensionLevel | 144 | dimensionLevelToAttribute | 144 |
| measureToSumAlongHie | 40 | sumAlongHieToMeasure | 40 |
| hierarchyToSumAlongHie | 611 | sumAlongHieToHierarchy | 611 |
| measureToSumAlongDim | 40 | sumAlongDimToMeasure | 40 |
| dimensionToSumAlongDim | 871 | sumAlongDimToDimension | 871 |
| measureToSumAlongFact | 40 | sumAlongFactToMeasure | 40 |
| factToSumAlongFact | 103 | sumAlongFactToFact | 103 |
| Total | 2.584 | Total | 2.584 |

Na Tabela 4.1 é apresentada a quantidade de instâncias incluídas no arquivo OWL para cada classe no metamodelo, totalizando 592 instâncias. Na Tabela 4.2 é apresentada a quantidade de relações mapeadas entre instâncias para cada relação entre

as classes do metamodelo totalizando 5.168 relações. As relações apresentadas na mesma linha da tabela são inversas entre si, por isso apresentam as mesmas quantidades. A definição das relações com suas correspondentes inversas na ontologia OWL possibilita a navegação entre as instâncias em qualquer direção em uma ferramenta de visualização.

Devido à grande quantidade de objetos na ontologia, o arquivo OWL resultante totalizou aproximadamente 36.00 linhas. Este tamanho elevado do arquivo ocasionou lentidão na visualização da ontologia através do Protégé. A ontologia OWL gerada pela execução do OntoDW está disponível em <https://sourceforge.net/projects/ontodw/files/estudo%20de%20caso/ontologia%20OWL>.

5 – Análise de Resultados

Este capítulo apresenta a análise dos resultados obtidos com o estudo de caso, através da avaliação do autor e da comparação com as pesquisas aplicadas.

5.1 Visão geral da análise dos resultados

Para uma avaliação mais abrangente das regras de mapeamento definidas e da utilidade da ontologia gerada pelo Onto DW através dos resultados obtidos com a execução do estudo de caso, foram realizadas:

- Análise pelo autor da ontologia gerada pelo OntoDW através da visualização do arquivo OWL pelo Protégé e do arquivo de LOG da execução, utilizando sua experiência de analista de BI e seu conhecimento do domínio e da organização do estudo de caso;
- 2 pesquisas, através de questionários, com especialistas da área de BI do mercado e analistas de TI com conhecimento e atividades relacionadas a BI da área de TI da organização do estudo de caso, para avaliação das regras de mapeamento definidas. Os resultados destas pesquisas foram comparados com o resultado gerado automaticamente com pelo OntoDW, que se propõe a realizar a tarefa de um analista de BI ao gerar um modelo conceitual;
- 1 pesquisa, através de reunião e questionário, com usuários do sistema objeto do estudo de caso, para avaliação da utilidade da ontologia gerada no apoio da tarefa de análise dos dados do DW;

Os resultados obtidos com estas avaliações estão detalhados nas subseções a seguir.

5.2 Análise da ontologia do estudo de caso

Para ilustrar os resultados obtidos com o estudo de caso, foi escolhido um trecho da ontologia gerada pela execução do OntoDW relacionado à agregabilidade de medi-

das ao longo de dimensões (**SummarizabilityAlongDimension**), por ele representar algumas das principais informações necessárias pelos usuários para a realização de análises, que são as dimensões pelas quais se pode analisar cada medida, e por esse trecho conter instâncias de diferentes classes extraídas. Ele contém elementos conceituais representando análises possíveis de analistas de negócio sobre os dados no Data Warehouse e apresenta diferentes conceitos extraídos.

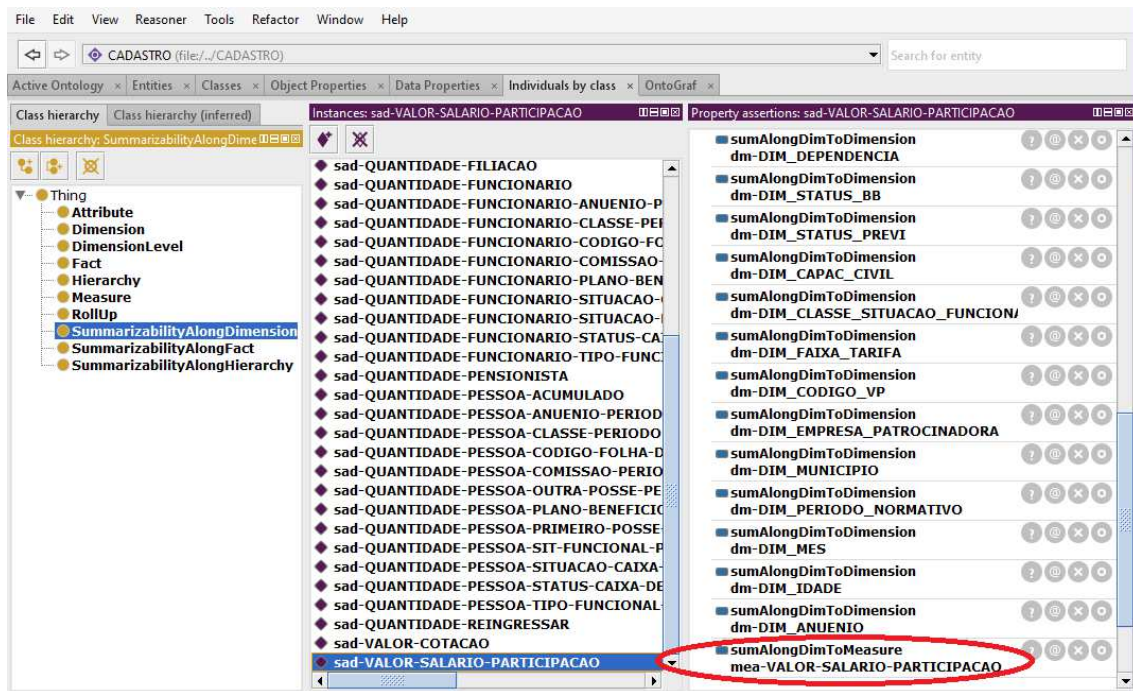


Figura 5.1: Captura de tela do Protégé com instâncias da ontologia

A Figura 5.1 ilustra uma captura de tela do Protégé apresentando instâncias da classe **SummarizabilityAlongDimension** encontradas pelo OntoDW. A tela ilustrada está dividida em três partes: na subdivisão mais à esquerda são apresentadas as classes definidas na ontologia (com uma das classes selecionada pelo usuário), na subdivisão central são apresentadas as instâncias da classe selecionada (com uma das instâncias selecionada) e na subdivisão mais à direita são apresentadas a métrica que tem sua agregabilidade representada (destacada em vermelho) e as dimensões pelas quais é possível analisar a métrica.

A métrica *mea-VALOR-SALARIO-PARTICIPACAO* destacada na Figura 5.1 possibilita a análise do salário dos funcionários (considerado pelo fundo de pensão para efeito de cálculo atuarial, pagamento de benefícios e arrecadação de obrigações dos participantes). O nome da métrica foi definido a partir do nome da coluna nas tabelas de

fato que armazena os seus dados (VAL_SALAR_PARTIC). Utilizando o separador (“_”) definido como parâmetro, os termos foram extraídos e consultados no glossário de termos de negócio da organização; se encontrado, o termo original é substituído pelo encontrado no glossário. Os termos são então concatenados com o outro separador (“-“), também definido como parâmetro. Um prefixo identificador da classe que define a instância também foi definido (“mea”), para auxiliar analistas de BI e analistas de negócio na identificação da classe a que a instância pertence.

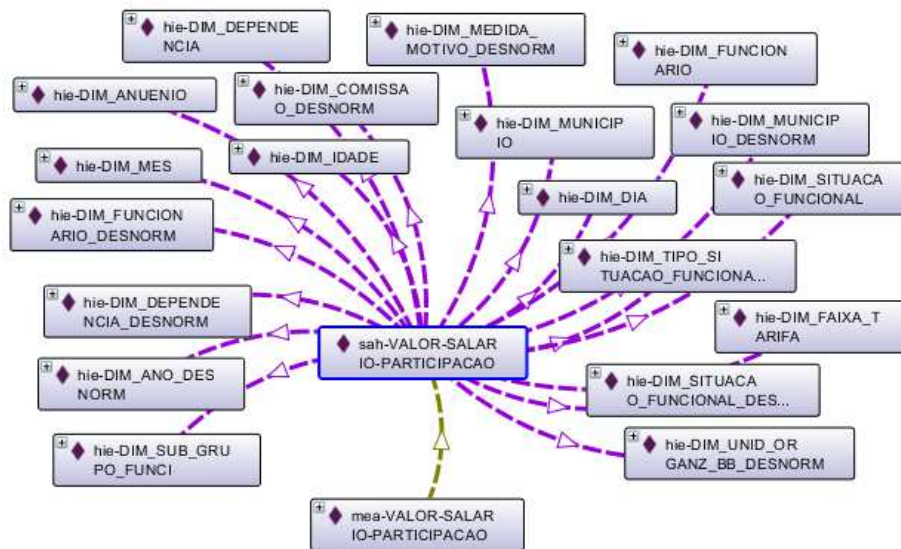


Figura 5.2: Recorte da ontologia extraída no estudo de caso

A instância com nome sad-VALOR-SALARIO-PARTICIPACAO (onde “sad” significa **SummarizabilityAlongDimension**) se relaciona com a medida mea-VALOR-SALARIO-PARTICIPACAO, conectando-as com todas as dimensões possíveis para análise. A Figura 5.2 é um recorte da mesma ontologia, representada graficamente através do plugin OntoGraf do Protégé (<https://github.com/protegeproject/ontograf>). Ele também apresenta a medida mea-VALOR-SALARIO-PARTICIPACAO, mas desta vez relacionada com a instância sah-VALOR-SALARIO-PARTICIPACAO (onde “sah” significa **SummarizabilityAlongHierarchy**), que a conecta com as hierarquias disponíveis para análise. As instâncias de hierarquia utilizam o prefixo identificador da classe (“hie”) em seus nomes. As instâncias da Figura 5.2 representam possibilidades de análise em uma maior granularidade que as instâncias da Figura 5.1.

Dentre os conceitos identificados pelo OntoDW, não foi encontrado nenhum incoerente com o DW em avaliação realizada pelo autor deste trabalho, exceto para as regras R1 e R3, como explicado a seguir.

A regra R1 tem o objetivo de identificar as tabelas de fato presentes no DW. O resultado obtido e não esperado foi a identificação da tabela REG_ULT_MES_CARGA como uma tabela do tipo fato sem fato. Esta tabela armazena dados relativos às cargas ocorridas no DW, e não informações de negócio. A identificação como tabela de fato ocorreu devido à sua estrutura; existe uma referência através de chave estrangeira a uma tabela de dimensão. A existência desta tabela é justificada pelo analista de BI pela necessidade de ter a informação de dia e hora da última carga em alguns relatórios, embora a inclusão desta informação no DW juntamente com os dados de negócio não seja usual. De acordo com o analista de BI, inclusive, na arquitetura do ambiente de BI da organização existe um banco de dados específico para as rotinas de ETL e tabelas de controle de carga. A identificação desta tabela como tabela de fato resultou na identificação de uma medida pela regra R6 para análise de quantidade de ocorrências entre as dimensões relacionadas a esta tabela. Apesar de inesperado e não aderente com os conceitos de negócio, esse resultado não pode ser considerado incorreto por representar uma análise possível de ser realizada com o esquema de dados. O resultado pode também servir para o analista de BI como um indicador de ajuste a ser feito no DW com a mudança de local da tabela, para torná-lo em conformidade com a arquitetura existente na organização.

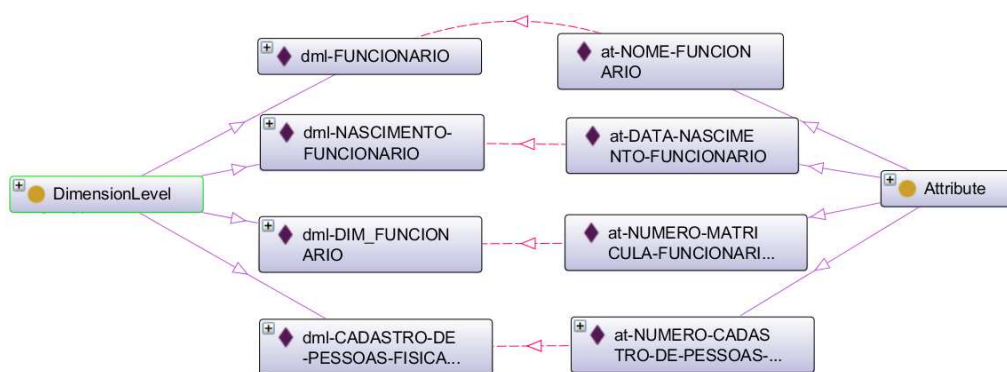


Figura 5.3: Níveis de dimensão da tabela DIM_FUNCIONARIO e atributos correspondentes

No caso do resultado inesperado obtido pela regra R3, foram identificados níveis de dimensão não esperados na tabela de dimensão DIM_FUNCIONARIO, que armazena

na os registros de funcionários e ex-funcionários. O resultado previsto era a existência de 1 só nível de dimensão (nível da chave primária, a menor granularidade da dimensão), com as colunas não associadas a alguma chave estrangeira ou não pertencentes à chave primária sendo qualificadores do funcionário. No entanto, foram extraídos 3 outros níveis de dimensão para 3 colunas da tabela que não são chaves estrangeiras e não pertencem à chave primária: NOM_FUNCI, DAT_NASC_FUNCI e NUM_CPF_FUNCI. Essas colunas podem ser vistas na representação da tabela existente DIM_FUNCIONARIO na Figura 4.1. O nível de dimensão esperado foi identificado somente com a coluna pertencente à chave primária como atributo (NUM_MATRIC_FUNCI). Estes 3 níveis adicionais foram encontrados devido à análise dos dados armazenados nas colunas da tabela. Nas 3 colunas foi constatada uma mesma ocorrência: alguns valores repetidos em registros com valores diferentes de chave primária, indicando que essas colunas podem ser utilizadas para agrupamento/consolidação dos dados e representam uma granularidade menor que o da chave primária. A coluna NOM_FUNCI armazena o nome do funcionário e a repetição de valores se justifica pela ocorrência de homônimos na tabela, comum pela quantidade de pessoas cadastradas. A coluna DAT_NASC_FUNCI armazena a data de nascimento do funcionário e a repetição de valores é normal neste caso. Por fim, a coluna NUM_CPF_FUNCI armazena o número de CPF das pessoas e a repetição de valores é justificada pelo fato da chave primária da tabela ser a matrícula do funcionário, e existirem casos em que um mesmo funcionário tem mais de uma matrícula associada a seu CPF, como os funcionários que se aposentam e posteriormente voltam a trabalhar na organização, por exemplo.

A Figura 5.3 apresenta alguns elementos da ontologia extraída onde são mostrados os níveis de dimensão identificados para a tabela DIM_FUNCIONARIO e seus respectivos atributos. Os nomes dessas instâncias foram definidos conforme o mesmo processo detalhado anteriormente para a definição de nome das medidas. O resultado encontrado na execução da regra R3 mostrou ser pertinente por explicitar análises possíveis de serem realizadas com os dados disponíveis na tabela do DW.

Considerando todos os conceitos automaticamente explicitados pelo OntoDW na ontologia gerada, o estudo de caso é considerado bem-sucedido. Com a ontologia graficamente representada, a identificação das instâncias e dos relacionamentos entre elas

pode ser feita de forma mais rápida e intuitiva. Retornando ao exemplo de análise sobre o impacto financeiro da aposentadoria de funcionários da instituição financeira, um analista de BI ou um usuário de negócio pode identificar a partir da Figura 5.2 as possibilidades de análise da métrica associada ao valor do salário em relação às dimensões encontradas no DW. Na Figura 5.3 é possível visualizar as subdivisões que a dimensão associada ao conceito Funcionário pode apresentar.

5.3 Análise das pesquisas com especialistas

Para mitigar o risco de uma análise enviesada da pesquisa, foram aplicados questionários online com especialista da área de Business Intelligence com o objetivo de avaliar a pertinência das regras de mapeamento. Os questionários online foram aplicados com a utilização da ferramenta google forms (<https://docs.google.com/forms>) e utilizando a escala Likert [Wuensch, 2005] para as questões objetivas relacionadas ao perfil do respondente. A escala Likert é um tipo de escala de resposta utilizada habitualmente em questionários, que solicita ao respondente seu nível de concordância com uma afirmação.

Conforme planejado, foram aplicados dois questionários semelhantes a dois públicos diferentes. O primeiro questionário foi direcionado a especialistas em BI em geral, sem necessária relação com a organização objeto do estudo de caso. Para abranger o maior número possível de profissionais, a pesquisa foi divulgada para: os contatos profissionais do pesquisador na área de BI (em torno de 70 pessoas), um grupo de integrantes do Programa de Pós-Graduação em Informática (PPGI) da UNIRIO (formado por professores, alunos e ex-alunos), e grupos do *LinkedIn* e *Facebook* relacionados a *Business Intelligence* e *Data Warehouse* no Brasil (com aproximadamente 8.000 inscritos no somatório de todos os grupos). O segundo questionário foi direcionado a profissionais que trabalham na organização objeto do estudo de caso com atividades associadas a sistemas de BI. O primeiro questionário aplicado se encontra no ANEXO I e o segundo questionário aplicado se encontra no ANEXO II.

A seguir são detalhadas as aplicações dos dois questionários, incluindo o número e o perfil dos respondentes, as questões apresentadas e a análise dos resultados. A análise dos resultados é qualitativa e busca identificar motivos para as respostas obtidas que divergem do resultado gerado pela execução das regras de mapeamento do OntoDW. Nesta análise são apresentadas as respostas objetivas de forma consolidada e pondera-

ções sobre as respostas subjetivas do segundo questionário. A totalidade das respostas subjetivas deste questionário se encontra no ANEXO IV.

5.3.1 Perfil dos respondentes

A primeira seção de perguntas nos dois questionários aplicados foi definida com o objetivo de analisar o perfil dos respondentes em relação ao conhecimento e experiência com modelagem multidimensional e aplicações OLAP.

Grupo 1 – Analistas de mercado especialistas em BI

O primeiro questionário, destinado a especialistas de mercado, foi respondido por 32 pessoas. Este grupo de pessoas também será chamado neste trabalho de grupo 1. A Figura 5.4 apresenta o grau de domínio do grupo 1 nos conceitos e na implementação de modelos multidimensionais. Ao analisar os gráficos, é possível considerar o grupo 1 como experiente em modelagem multidimensional pelo fato da maioria das pessoas se enquadrarem nas opções 4 ou 5. Entretanto, a concentração de pessoas nessas opções é maior para o domínio nos conceitos, com 78,2% dos respondentes, enquanto que para o domínio na implementação de modelos a concentração de respostas nas opções 4 e 5 foi de 52,5%. Isso mostra uma maior perícia do grupo com teoria em relação à prática em modelagem multidimensional.

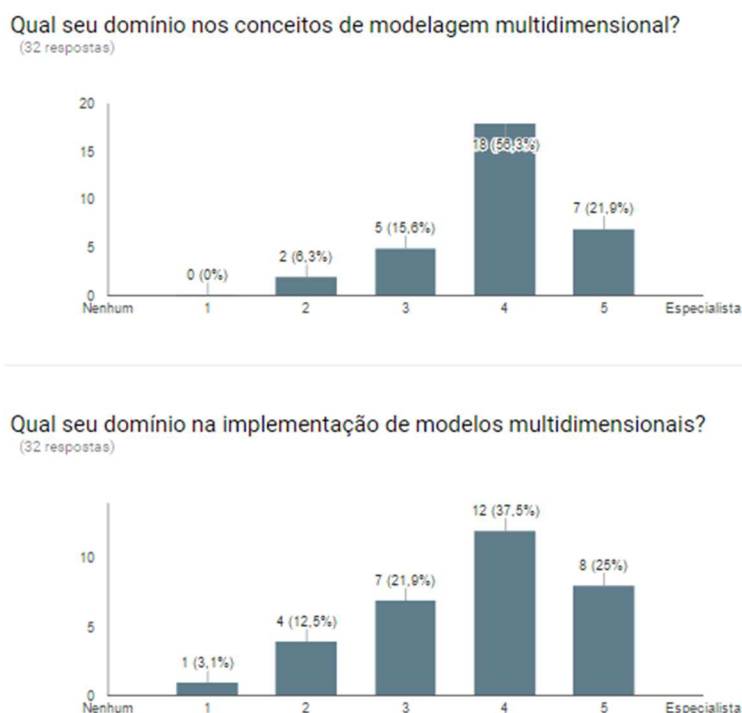
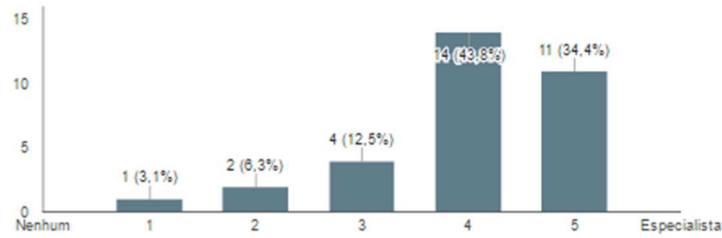


Figura 5.4: Perfil do grupo 1 em relação a modelagem multidimensional

Qual seu domínio nos conceitos de aplicações OLAP (ex.: roll up, hierarquia, atributo, métrica)?
(32 respostas)



Qual seu domínio na implementação de aplicações OLAP? (32 respostas)

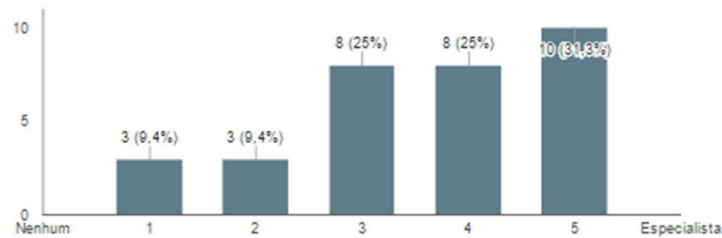


Figura 5.5: Perfil do grupo 1 em relação a aplicações OLAP

A Figura 5.5 apresenta o grau de domínio do grupo 1 nos conceitos e na implementação de aplicações OLAP. Ao analisar os gráficos, também é possível considerar o grupo como experiente em aplicações OLAP pela maioria das respostas nas opções 4 ou 5. Também houve uma concentração maior nessas opções para o domínio nos conceitos, com 78,2% dos respondentes, enquanto que para o domínio na implementação a concentração de respostas nas opções 4 e 5 foi de 56,3%. Isso mostra uma maior perícia do grupo com teoria em relação à prática também em aplicações OLAP.

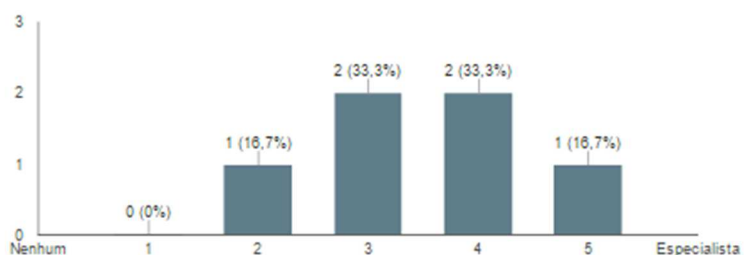


Figura 5.6: Perfil do grupo 1 em relação à sua experiência

A Figura 5.6 apresenta o tempo de experiência em anos dos respondentes do grupo 1 na implementação de modelos multidimensionais (gráfico à esquerda) e implementação de aplicações OLAP (gráfico à direita). Esses gráficos reforçam a definição desse grupo de respondentes como experiente, pois 84,3% das pessoas têm pelo menos 2 anos de experiência na implementação de modelos multidimensionais e 75% das pessoas têm pelo menos 2 anos de experiência na implementação de aplicações OLAP.

Grupo 2 – Analistas de TI da organização do estudo de caso

Qual seu domínio nos conceitos de modelagem multidimensional? (6 respostas)



Qual seu domínio na implementação de modelos multidimensionais? (6 respostas)

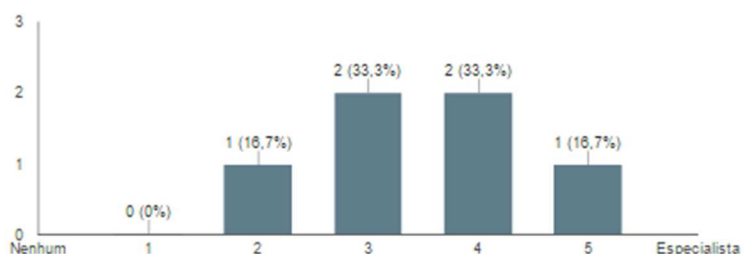
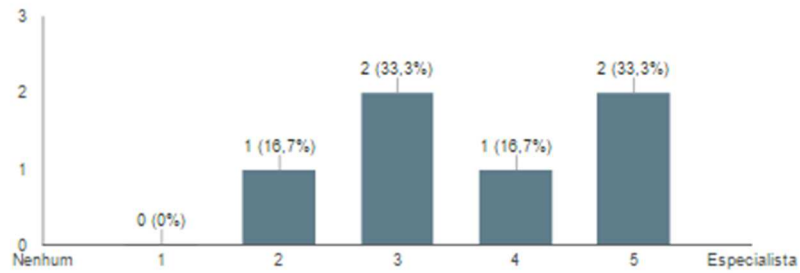


Figura 5.7: Perfil do grupo 2 em relação a modelagem multidimensional

O segundo questionário, destinado a especialistas de TI da organização do estudo de caso, foi respondido por 6 pessoas. Este grupo de pessoas também será chamado neste trabalho de grupo 2. A Figura 5.7 apresenta o grau de domínio dos respondentes do grupo 2 nos conceitos e na implementação de modelos multidimensionais e a Figura 5.8 seu grau de domínio nos conceitos e na implementação de aplicações OLAP. Para este grupo, metade dos respondentes se enquadraram nas opções 4 ou 5 para todas as questões, apresentando o mesmo domínio para teoria e prática. Isso mostra uma menor perícia deste grupo de respondentes, composto por analistas de TI da organização, em relação aos respondentes do questionário 1, composto por especialistas em BI. No en-

tanto, os respondentes comprovadamente trabalham com atividades associadas a sistemas de BI e têm informações adicionais por trabalharem na organização, o que juntamente com o perfil do grupo os torna aptos a responder ao questionário.

Qual seu domínio nos conceitos de aplicações OLAP (ex.: roll up, hierarquia, atributo, métrica)?
(6 respostas)



Qual seu domínio na implementação de aplicações OLAP? (6 respostas)

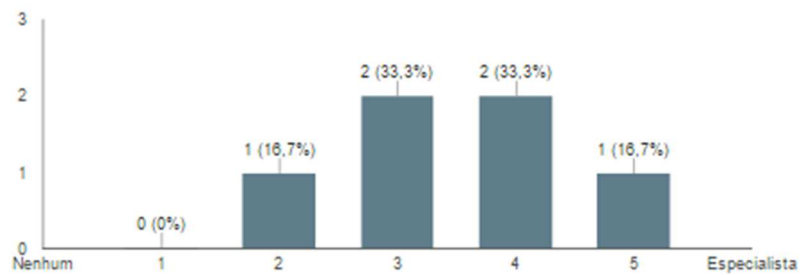
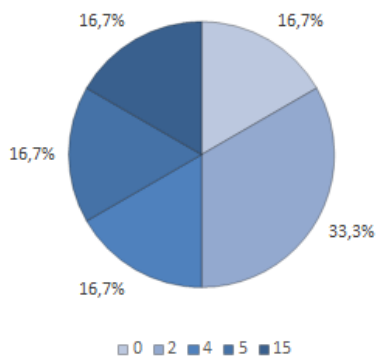


Figura 5.8: Perfil do grupo 2 em relação a aplicações OLAP

Qual sua experiência, em anos, na implementação de modelos multidimensionais?



Qual sua experiência, em anos, na implementação de aplicações OLAP?

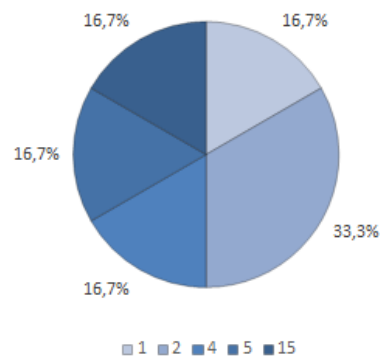


Figura 5.9: Perfil do grupo 2 em relação à sua experiência

A Figura 5.9 apresenta o tempo de experiência em anos dos respondentes do questionário 2 na implementação de modelos multidimensionais (gráfico à esquerda) e implementação de aplicações OLAP (gráfico à direita). Esses gráficos demonstram que este grupo de respondentes também é experiente, pois 83,3% das pessoas têm pelo menos 2 anos de experiência na implementação de modelos multidimensionais e na implementação de aplicações OLAP. Como esse grupo é bem menor que o grupo do questionário 1 (6 respondentes contra 32 respondentes, respectivamente), é de fácil percepção uma maior distribuição da experiência das pessoas. Enquanto uma pessoa se declarou como sem experiência em um gráfico, outra informou ter 15 anos de experiência.



Figura 5.10: Perfil do grupo 2 em relação a tempo de empresa

Para o grupo 2 de respondentes foram apresentadas duas perguntas adicionais relacionadas ao seu perfil. O objetivo dessas perguntas era obter mais informações sobre o nível de conhecimento a respeito de aspectos mais específicos à organização onde trabalham através de questões relacionadas ao vínculo de trabalho das pessoas. A primeira pergunta foi sobre o tempo que trabalham na organização. Uma consolidação das respostas obtidas está apresentada na Figura 10 e mostra um grupo familiarizado com a organização, com 66,6% das pessoas com pelo menos 6 anos de trabalho. A segunda questão solicitava a função ocupada pelo respondente. Todos informaram realizar funções associadas de alguma forma com sistemas de BI. O grupo incluiu 1 gerente de desenvolvimento de sistemas, 1 arquiteto de BI, 2 administradores de dados e 2 analistas de sistemas.

5.3.2 Resultados das regras R1 e R2

A segunda seção de perguntas dos dois primeiros questionários aplicados foi definida com o objetivo de avaliar as regras R1 e R2 do OntoDW. A regra R1 busca identificar instâncias da classe *Fact* e a regra R2 identificar instâncias da classe *Dimension*.

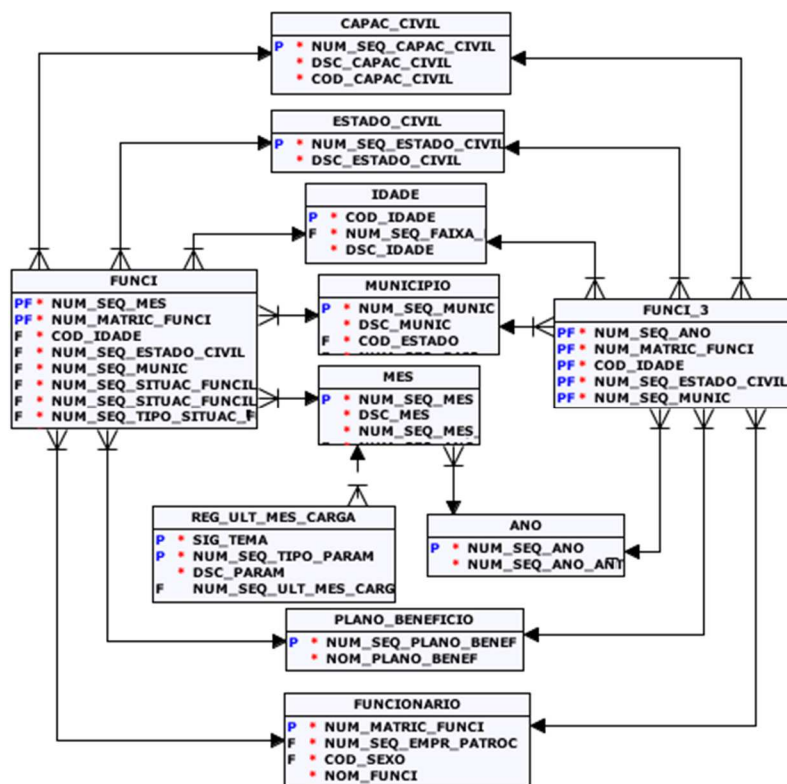


Figura 5.11: Recorte 1 do DW utilizado para validar as regras R1 e R2

O questionário incluiu o recorte do modelo do DW ilustrado na Figura 5.11 e as questões apresentadas nas Figuras 5.12 e 5.13. O modelo da Figura 5.11 é semelhante ao modelo de dados ilustrado na Figura 4.2. A diferença é que na Figura 5.11 os nomes das tabelas tiveram seu prefixo retirado, pois sugeriam qual a função da tabela, se era fato ou dimensão. Em todos os recortes de modelo incluídos nos questionários foram retirados os prefixos dos nomes das tabelas, por esse mesmo motivo.

Para colher as respostas dos grupos, foi formulada a questão apresentada na Figura 5.12. Primeiro, para cada tabela do recorte, o respondente indicava se a mesma seria uma tabela de dimensão, uma tabela de fato/agregação ou nenhuma das duas opções. Além disso, as questões subjetivas da Figura 5.13 foram feitas para obter a opinião

dos respondentes do grupo 2 a respeito do recorte de modelo apresentado e das impressões ao responder a pesquisa. Essas mesmas questões subjetivas foram incluídas a cada novo recorte de modelo mostrado, que ocorre a cada mudança de seção do formulário. A primeira questão sofre uma pequena alteração ao ser reproduzida, com o ajuste do nome do modelo para aquele presente na seção.

01. Informe qual conceito de modelagem dimensional (dimensão ou fato/agregação) corresponde a cada tabela do Recorte 1, se for o caso: *

| | DIMENSÃO | FATO/AGREGAÇÃO | NENHUM DOS DOIS |
|-------------------|-----------------------|-----------------------|-----------------------|
| ANO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| CAPAC_CIVIL | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| ESTADO_CIVIL | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCI | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCIONARIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCI_3 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| IDADE | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| MES | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| MUNICIPIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| PLANO_BENEFICIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| REG_ULT_MES_CARGA | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Figura 5.12: Questão para validar as regras R1 e R2

As informações presentes no Recorte 1 foram suficientes para responder a questão acima? De qual informação sentiu falta?

Sua resposta

Comente sobre a dificuldade de responder à questão acima, a estratégia para chegar à resposta ou qualquer outro assunto que achar pertinente.

Sua resposta

Figura 5.13: Questões subjetivas sobre a seção 4 do formulário para o grupo 2

As respostas foram agrupadas conforme o resultado obtido pela execução do OntoDW. Dessa forma, as tabelas foram divididas entre as que foram mapeadas como fatos

pela regra R1 e as que foram mapeadas como dimensões pela regra R2. Após a separação, foram analisadas as respostas dos 2 grupos de respondentes.

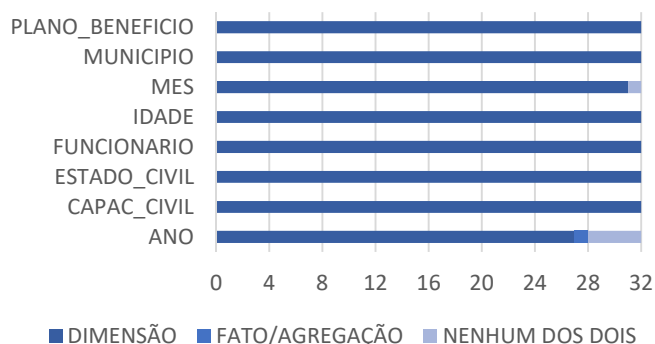


Figura 5.14: Respostas do grupo 1 para as tabelas de dimensão identificadas pelo OntoDW através da regra R1

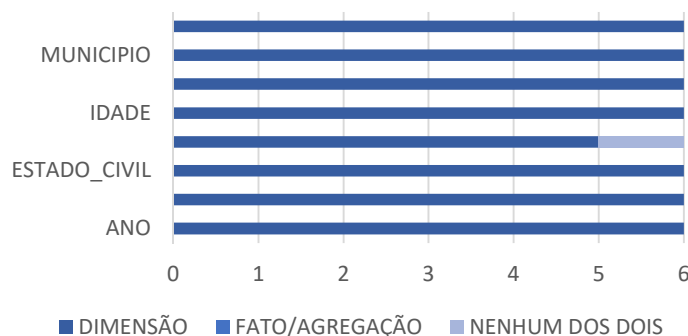


Figura 5.15: Respostas do grupo 2 para as tabelas de dimensão identificadas pelo OntoDW através da regra R1

A Figura 5.14 apresenta a consolidação das respostas do grupo 1 para as tabelas mapeadas pelo OntoDW como dimensões, que foram 8 dentre as apresentadas na Figura 5.11. Dessas 8 tabelas, 6 foram identificadas por todos os respondentes como dimensões. Das duas tabelas restantes, a tabela MES foi marcada como dimensão por 31 pessoas (96,9% do total) e por 1 pessoa como “NENHUM DOS DOIS”. A tabela ANO foi marcada como dimensão por 27 pessoas (84,4% do total), como fato/agregação por 1 pessoa e como nenhuma dessas duas opções por outras 4 pessoas. Foi identificado que a única pessoa que não identificou a tabela MES como dimensão também foi uma das pessoas que também não identificou a tabela ANO como dimensão. Ela marcou a opção “NENHUM DOS DOIS” para ambos.

Analisando o perfil dos respondentes não foram encontrados padrões ou características que justificassem as respostas que diferiam da maioria do grupo. Entretanto, as diferenças ocorreram nas duas dimensões com relacionamentos diferentes das demais. A tabela MES está relacionada com a tabela REG_ULT_MES_CARGA, que o nome sugere que ela armazene informações de controle. A tabela ANO é a única referenciada por outra dimensão. As outras dimensões apresentam as mesmas relações. Isso sugere que a tabela REG_ULT_MES_CARGA e a relação existente entre as 2 dimensões tornou confusa a análise do modelo para essas pessoas e que a nomenclatura da tabela influenciou nas respostas, mesmo com a retirada dos prefixos.

A Figura 5.15 apresenta a consolidação das respostas do grupo 2 para as tabelas mapeadas pelo OntoDW como dimensões. Neste grupo, das 8 tabelas, 7 foram identificadas por todos os respondentes como dimensões. A tabela FUNCIONARIO foi marcada como dimensão por 7 pessoas (87,5% do total) e por 1 pessoa como “NENHUM DOS DOIS”. Esta pessoa se apresenta pelo seu perfil como a menos qualificada do grupo e com menos de 1 ano de experiência na implementação de modelos dimensionais. Algum aspecto do modelo pode tê-la confundido.

Mesmo sem atingir 100% de acerto nas respostas nos dois grupos, pode-se admitir como bem-sucedida a regra R1 pelo alto índice de acerto atingido nas respostas dos grupos e pelas considerações descritas acima. Dessa forma, conseguiu-se explicitar um conhecimento difundido entre analistas de BI no sentido de definir critérios para a realização de análises uniformes de modelos multidimensionais de dados.

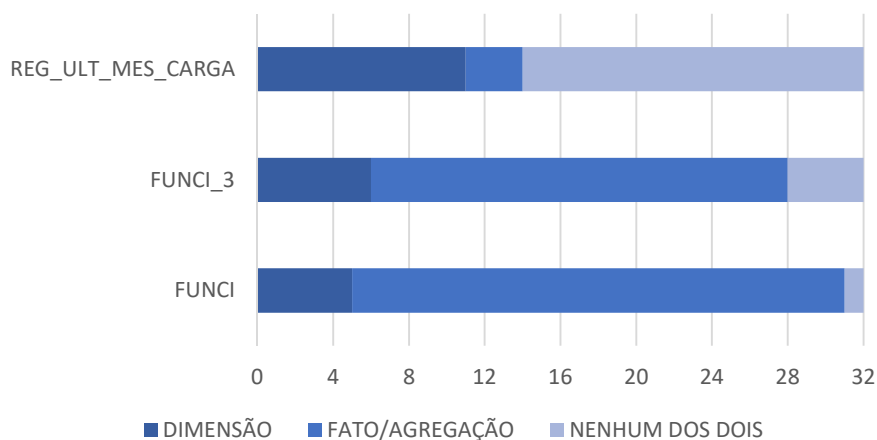


Figura 5.16: Respostas do grupo 1 para as tabelas de fato identificadas pelo OntoDW através da regra R2

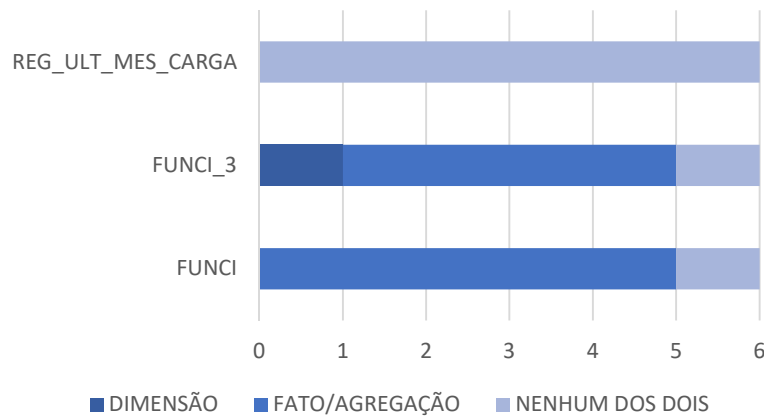


Figura 5.17: Respostas do grupo 2 para as tabelas de fato identificadas pelo OntoDW através da regra R2

A Figura 5.16 apresenta a consolidação das respostas do grupo 1 para as tabelas mapeadas pelo OntoDW como fatos (as 3 tabelas restantes da Figura 5.11). Nenhuma das tabelas foi identificada como fato por todo o grupo. A tabela REG_ULT_MES_CARGA foi marcada como fato por 3 pessoas (9,38% do total), a tabela FUNC_3 foi marcada como fato por 22 pessoas (68,75% do total) e a tabela FUNC_1 foi marcada como por 26 pessoas (81,25% do total).

Conforme descrito na Seção 5.1, a tabela REG_ULT_MES_CARGA não é utilizada como tabela de fato no DW estudado, embora sua identificação pelo OntoDW tenha sido considerada coerente de acordo com a regra. A estrutura da tabela possibilita que a mesma seja utilizada como fato, mas sua utilização não fica explícita. Assim, duas justificativas para as respostas quanto à essa tabela podem ser o nome da tabela, que não deixa claro sua utilidade, e as colunas que são apresentadas. Uma das colunas tem o nome iniciado com o prefixo “DSC”, sugerindo que seja uma coluna de descrição, mas usada em dimensões.

Quanto às outras duas tabelas, a maioria das respostas na Figura 5.16 as qualificou como fato/agregação. Para estas tabelas, duas justificativas para estas respostas são a falta de nomes mais explícitos das tabelas, sugerindo que as mesmas são fatos, e a ausência de colunas sem chave estrangeira e fora da chave primária, o que indicaria a presença de medidas/métricas.

A Figura 5.17 apresenta a consolidação das respostas do grupo 2 para as tabelas mapeadas pelo OntoDW como fatos. Neste grupo, a tabela REG_ULT_MES_CARGA

foi identificada por todos como “NENHUM DOS DOIS”. Isto mostra que o conhecimento dos padrões da organização para a definição dos objetos é importante na análise dos modelos. A tabela FUNCI_3 foi identificada por 4 respondentes (66,67% do total) como fato/agregação e a tabela FUNCI foi identificada por 5 respondentes (83,33% do total) como fato/agregação. Uma consideração relevante é que o respondente menos qualificado do grupo 2 não identificou nenhuma das duas tabelas como fato. A tabela FUNCI_3 também recebeu uma qualificação como dimensão de um respondente.

Para essas respostas divergentes do OntoDW, podem ser atribuídas possíveis dificuldades dos respondentes por falta de experiência ou pela falta de nomenclatura e formatação explícita da tabela. Isso devido às informações prestadas pelos respondentes nas questões subjetivas. Metade das pessoas informou que utilizou como estratégia a análise das relações entre as tabelas para definir a resposta. Também metade das pessoas informou que a ausência de um nome mais semântico para as tabelas, como um prefixo indicando seu uso, dificultou a análise do modelo. Houve ainda uma sugestão de organização das tabelas, com os fatos no centro do modelo e as dimensões nas bordas ou ainda com as tabelas coloridas por sua utilidade, para facilitar uma identificação visual da função das tabelas.

A regra R2 foi considerada parcialmente bem-sucedida. Foi positivo o alto índice de acerto atingido nas respostas dos grupos para as tabelas FUNCI e FUNCI_3, além das considerações descritas acima. Entretanto, para a tabela REG_ULT_MES_CARGA foi obtida mais de uma interpretação de sua utilidade pelos respondentes. Apesar de a mesma apresentar estrutura coerente com uma tabela de fato, ela não é utilizada no DW com este fim. Assim, os critérios definidos na regra R2 produziram um resultado incoerente com o uso real no DW para esta tabela de controle. Como o uso do conhecimento dos padrões da organização para os sistemas se mostrou importante, uma sugestão é refinar a regra R2 para que utilize informações do metamodelo de domínio na sua definição.

5.3.3 Resultados das regras R3, R4, R8 e R10

A terceira seção de perguntas foi definida para avaliar as regras R3, R4, R8 e R10 do OntoDW. A regra R3 busca identificar instâncias da classe **DimensionLevel** e a regra R4 as instâncias da classe **Attribute**. As regras R8 e R10 buscam identificar as

instâncias das classes **RollUp** e **Hierarchy**, respectivamente, mas analisando os níveis de 1 só dimensão. As instâncias dessas classes envolvendo mais de uma dimensão são identificadas por outras regras.

| FUNCIONARIO ▲ | |
|---------------|--------------------|
| P | NUM_MATRIC_FUNC1 |
| F | COD_SEXO |
| | NOM_FUNC1 |
| | DAT_NASC_FUNC1 |
| | NUM_CPF_FUNC1 |
| F | NUM_SEQ_DIA_APOSE |
| F | NUM_SEQ_DIA_EXONDO |
| F | NUM_SEQ_DIA_FALECI |

Figura 5.18: Recorte 2 do DW utilizado para validar as regras R3, R4, R8 e R10

02. Considerando que a tabela FUNCIONARIO apresentada no Recorte 2 é uma dimensão, informe o número de níveis de dimensão que ela contém: *

Sua resposta

03. Baseado nos níveis de dimensão encontrados na tabela do Recorte 2, informe o número de operações de Roll Up (mudança de grão para um nível hierarquicamente superior) que podem ser realizadas: *

Sua resposta

04. Caso tenha encontrado Roll Up's na tabela do Recorte 2, eles poderiam ser agrupados em uma hierarquia?: *

Sim

Não

05. Cada coluna da matriz abaixo representa um possível nível de dimensão encontrado. Agrupe atributos (colunas) que sejam de um mesmo nível de dimensão a uma mesma coluna. Ex.: Caso tenha encontrado apenas 1 nível de dimensão, marque a 1ª coluna em todas as linhas: *

Atributos (Colunas da tabela) x Níveis de dimensão

| | Nível 1 | Nível 2 | Nível 3 | Nível 4 | Nível 5 | Nível 6 |
|--------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| NUM_MATRIC_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| COD_SEXO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NOM_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| DAT_NASC_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_CPF_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_APOSE | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_EXONDO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_FALECI | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Figura 5.19: Questões para validar as regras R3, R8, R10 e R4

As informações presentes no Recorte 2 foram suficientes para responder às questões acima? De qual informação sentiu falta?

Sua resposta

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

Sua resposta

Figura 5.20: Questões subjetivas sobre a seção 5 do formulário para o grupo 2

O questionário incluiu o recorte do modelo do DW ilustrado na Figura 5.18 com a tabela FUNCIONARIO e as questões apresentadas na Figura 5.19. Além disso, as questões subjetivas da Figura 5.20 foram feitas para obter a opinião do grupo 2 a respeito

to desta seção do formulário. As respostas foram agrupadas para comparação com o resultado obtido pela execução do OntoDW.

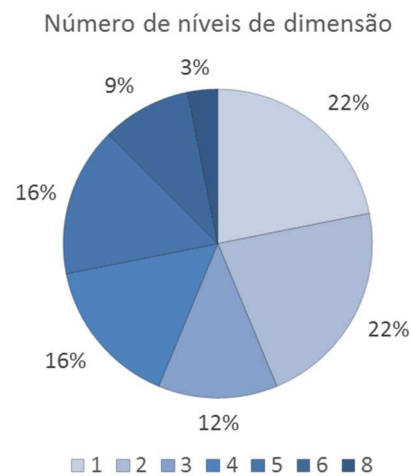


Figura 5.21: Respostas do grupo 1 para a questão 02

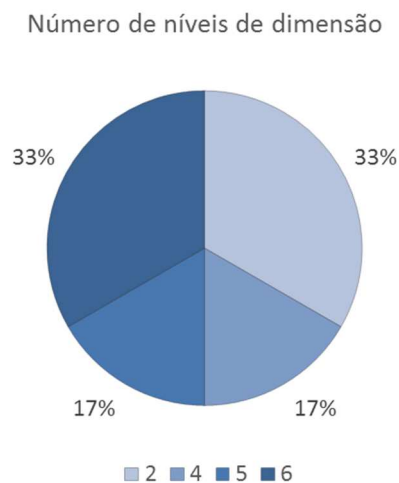


Figura 5.22: Respostas do grupo 2 para a questão 02

A questão 02 informa que a tabela FUNCIONARIO é uma tabela de dimensão e solicita que seja informado o número de níveis de dimensão encontrados. Ela objetiva analisar o resultado da regra R3. Na Figura 5.21 e na Figura 5.22 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2 para a questão 02, respectivamente. É possível perceber nas respostas dos dois grupos que não houve uma quantidade de níveis com maioria de escolha, com percepções diferentes entre as pessoas.

Tabela 5.1: Níveis de dimensão da tabela FUNCIONARIO

| |
|---|
| Níveis de dimensão |
| DIM_FUNCIONARIO |
| FUNCIONARIO |
| CADASTRO-DE-PESSOAS-FISICAS-FUNCIONARIO |
| NASCIMENTO-FUNCIONARIO |

Conforme demonstrado na Figura 5.3, foram encontrados 4 níveis para esta dimensão pelo OntoDW, apesar de a expectativa é de que apenas 1 fosse encontrado. Estes níveis podem ser vistos na Tabela 5.1. Os nomes apresentados para eles foram definidos a partir de consulta ao glossário da organização com os termos extraídos das colunas da tabela.

Ficou evidente a existência de múltiplos critérios para a contagem do número de níveis pelos respondentes. Pela análise dos comentários do grupo 2 (colhidos através das questões da Figura 5.20) foram citados:

- Dificuldade no entendimento do conceito de níveis de dimensão;
- Tentativa de identificação de níveis através das chaves estrangeiras, fazendo com que fossem contabilizados os níveis de outras tabelas de dimensão;
- Tentativa de identificação de colunas que qualificassem um funcionário para agrupá-las.

Nenhum participante do grupo 2 comentou ter consultado os dados da tabela, apesar dessa possibilidade existir. Analisando o perfil dos respondentes não foram encontrados padrões ou características que motivassem os agrupamentos de respostas encontrados, apenas comentários de dificuldade ao responder as questões pelos analistas menos experientes do grupo 2.

Podemos concluir, com as respostas consolidadas e com o próprio resultado obtido pelo OntoDW, que a tarefa de identificação de níveis de uma dimensão não é trivial e se mostrou muito dependente da interpretação do analista. Outra constatação foi que a utilização dos dados da tabela se mostrou essencial para explicitar relações existentes na prática, mas que teoricamente não são consideradas, como o agrupamento de pessoas homônimas, por exemplo. Podemos definir a regra R3 como pertinente por gerar um resultado coerente, além de que os resultados obtidos pelas pesquisas mostram uma falta de uniformidade de conceito pelos analistas. Essa falta de consenso explicita a necessi-

dade de iniciativas para a de definição de critérios para prover a realização de análises mais uniformes de modelos multidimensionais por diferentes analistas.

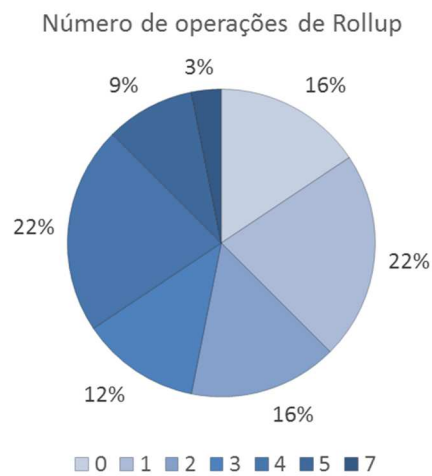


Figura 5.23: Respostas do grupo 1 para a questão 03

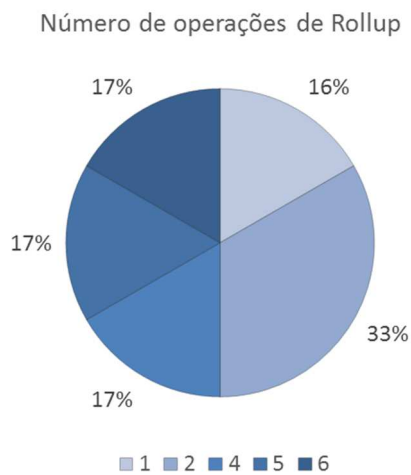


Figura 5.24: Respostas do grupo 2 para a questão 03

A questão 03 solicita que seja informado o número de operações de *roll up* possíveis de acordo com os níveis de dimensão encontrados na tabela e informado na questão 02. Essa questão objetiva analisar o resultado da regra R8. Na Figura 5.23 e na Figura 5.24 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2 para a questão 03, respectivamente.

Tabela 5.2: Operações de *roll up* da dimensão FUNCIONARIO

| Nível de origem | Nível de destino |
|---|---|
| DIM_FUNCIONARIO | FUNCIONARIO |
| FUNCIONARIO | CADASTRO-DE-PESSOAS-FISICAS-FUNCIONARIO |
| CADASTRO-DE-PESSOAS-FISICAS-FUNCIONARIO | NASCIMENTO-FUNCIONARIO |

A execução do OntoDW encontrou 3 operações de *roll up* possíveis para a tabela FUNCIONARIO, listadas na Tabela 5.2. As operações possíveis foram definidas a partir dos níveis encontrados para esta tabela e listados na Tabela 5.1.

Analisando o resultado obtido com a questão 03, não houve uma quantidade de operações com maioria de escolha e o resultado foi bem diverso. Esse resultado é esperado pela dependência da identificação dos níveis de dimensão para identificação das operações possíveis de *roll up*. Nos comentários fornecidos pelo grupo 2, foi identificada a tentativa de encontrar as operações de *roll up* através da análise das chaves estrangeiras. Ficou claro que as colunas com chave estrangeira foram contabilizadas por algumas pessoas pelo alto número de operações informado por algumas pessoas. O OntoDW trata as colunas com chave estrangeira apenas como relações entre tabelas.

Os números encontrados para as operações de *roll up* se mostraram alinhados com os níveis de dimensão encontrados. No entanto, fica ressaltado que o número alto de níveis de dimensão encontrados por algumas pessoas está relacionado às chaves estrangeiras. Caso sejam consideradas as colunas com chave estrangeira como nível de dimensão, os níveis de dimensão serão contados em dobro, pois já são contabilizados nas tabelas que os contém, que são referenciadas pela tabela FUNCIONARIO.

A tarefa de identificação das operações possíveis de *roll up* também se mostrou muito dependente da interpretação do analista e da correta identificação dos níveis da dimensão. Podemos definir a regra R8 como bem-sucedida pelo número correto de operações encontrado estar correto; porém, o nível de origem definido para as operações pode ser passível de questionamentos. Pela análise dos dados e metadados, o OntoDW identificou, por exemplo, que o nível que representa o nome do funcionário (FUNCIONARIO) é hierarquicamente inferior ao nível que representa o CPF do funcionário (CADASTRO-DE-PESSOAS-FISICAS-FUNCIONARIO). Nos termos do negócio, o CPF não representaria repetição, estando na menor granularidade possível, e o

nome do funcionário poderia apresentar repetição. No entanto, vale ressaltar que as operações da Tabela 5.2 extraídas pelo OntoDW estão condizentes com o estado do DW.

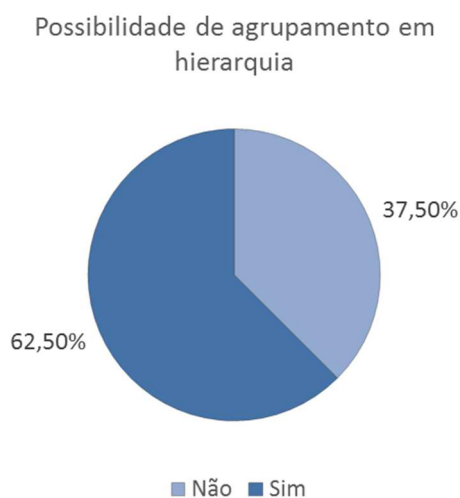


Figura 5.25: Respostas do grupo 1 para a questão 04

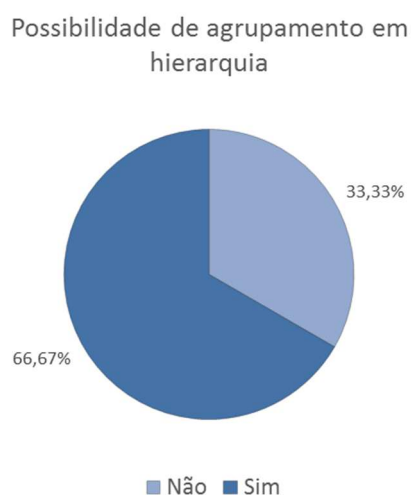


Figura 5.26: Respostas do grupo 2 para a questão 04

A questão 04 solicita que seja informado se, caso o respondente tenha encontrado *roll up's* possíveis na dimensão FUNCIONARIO, eles podem ser agrupados em uma hierarquia. Essa questão objetiva analisar o resultado da regra R10. Na Figura 5.25 e na Figura 5.26 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2 para a questão 04, respectivamente.

É possível observar que o resultado para os dois grupos foi muito próximo, com a maioria das pessoas (em torno de 60%) informando que é possível agrupar os *roll up's* encontrados em uma hierarquia. Nos comentários fornecidos pelo grupo 2, as duas pes-

soas que responderam “Não” afirmaram que as chaves estrangeiras representavam um segundo nível em direções diferentes em relação à dimensão FUNCIONARIO. Mais uma vez, as chaves estrangeiras foram importantes na definição das respostas. Devemos considerar também que 7 pessoas no grupo 1 e 1 pessoa no grupo 2 informaram que encontram um número menor que 2 *roll up*'s na dimensão. É possível que não tenham identificado a necessidade de criar uma hierarquia com 1 ou nenhum *roll up*. Parte dessas pessoas poderia ter respondido diferente ao encontrar um número maior de *roll up*'s, aumentando o percentual de respostas “Sim”.

A tarefa de identificação da hierarquia é fortemente relacionada com a identificação correta dos *roll up*'s. Caso isso seja feito, identificar a hierarquia fica relativamente simples. Podemos definir a regra R10 como bem-sucedida por gerar um resultado correto.

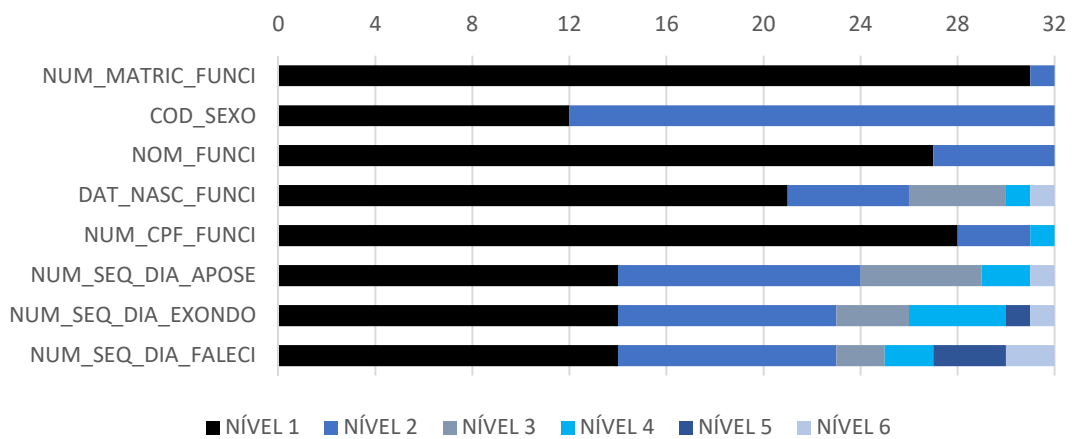


Figura 5.27: Respostas do grupo 1 para a questão 05

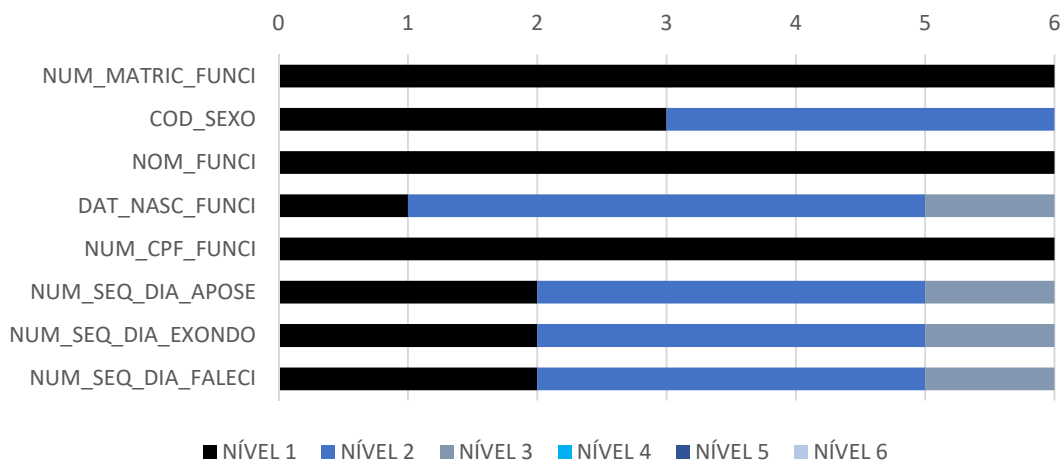


Figura 5.28: Respostas do grupo 2 para a questão 05

A questão 05 solicita que sejam vinculados os atributos (colunas) aos níveis encontrados da dimensão FUNCIONARIO. Essa questão objetiva analisar o resultado da regra R4. Na Figura 5.27 e na Figura 5.28 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2 para a questão 05, respectivamente.

Tabela 5.3: Níveis da dimensão FUNCIONARIO e atributos correspondentes

| Nível de dimensão | Atributo |
|---|-----------------|
| DIM_FUNCIONARIO | NUM_MATRIC_FUNC |
| FUNCIONARIO | NOM_FUNC |
| NASCIMENTO-FUNCIONARIO | DAT_NASC_FUNC |
| CADASTRO-DE-PESSOAS-FISICAS-FUNCIONARIO | NUM_CPF_FUNC |

Na tabela FUNCIONARIO, a execução do OntoDW encontrou 1 atributo para cada um dos 4 níveis de dimensão anteriormente definidos. Os atributos encontrados e seus níveis correspondentes estão apresentados na Tabela 5.3. As colunas que possuem chave estrangeira não são identificadas como atributos.

No grupo 1, o resultado encontrado se mostrou bem diferente do resultado fornecido pelo OntoDW, mas semelhante ao que era previsto para o estudo de caso antes de sua aplicação. Conforme detalhado na Seção 5.1, era esperado apenas 1 nível para a dimensão FUNCIONARIO. Assim, todos os atributos estariam vinculados a esse nível único, da mesma granularidade da chave primária. Para este grupo, a maioria das pessoas vinculou os atributos listados na Tabela 5.3 ao Nível 1, agrupando-os num mesmo nível.

É provável que a semântica presente no nome da coluna foi utilizada pelos respondentes para esta atividade, assim como a estrutura da tabela. A coluna com maior vinculação ao nível 1 foi a NUM_MATRIC_FUNC (96,9% dos respondentes), que pertence à chave primária e seu nome sugere que represente a matrícula do funcionário, comumente encontrada como identificador em organizações. A segunda maior foi a NUM_CPF_FUNC (87,5% dos respondentes), que não pertence à chave primária, mas seu nome indica que representa o CPF do funcionário, comumente utilizado como identificador de indivíduos. A terceira maior foi a NOM_FUNC (84,4% dos respondentes), que claramente é um qualificador de um indivíduo, apesar da existência de homônimos. Por último foi a DAT_NASC_FUNC (65,6% dos respondentes), que também é um

qualificador de um indivíduo, mas é mais evidente a possibilidade de repetição de valores para mais de um registro.

Comportamento semelhante aconteceu com o grupo 2 para 3 dentre os 4 atributos. A exceção foi o DAT_NASC_FUNCI, que apenas 1 pessoa (ou 16,7% dos respondentes) vinculou ao mesmo nível dos outros atributos, justamente o analista menos experiente do grupo. Isso sugere que um maior conhecimento do domínio e dos sistemas da organização indicaram à maioria do grupo que esta coluna pertence a um nível com menor granularidade que o da chave primária. Para os 3 atributos, 100% deste grupo agrupou-os no nível 1.

Em ambos os grupos, as colunas com chave estrangeira tiveram na maioria de suas respostas uma opção diferente do nível 1. Alguns respondentes vincularam cada uma dessas colunas a um nível diferente e outros respondentes agruparam tais colunas num só nível. Como o grupo 1 é composto de mais pessoas e tem um perfil mais diverso (todos do grupo 2 trabalham na mesma organização), seu gráfico (Figura 5.27) apresentou respostas com maior distribuição entre os níveis para as colunas com chave estrangeira que o gráfico de respostas do grupo 2 (Figura 5.28).

Para a regra R4, o resultado colhido nas pesquisas foi bem diferente do produzido pelo OntoDW. Isso ressaltou a dificuldade de identificar o conhecimento que está implícito nas estruturas de dados pelos analistas. Em casos em que existia semântica explícita nos nomes das colunas, o resultado foi mais próximo de um consenso. Como exemplo, pode-se observar que mais de 65% dos respondentes do grupo 1 agruparam as 4 colunas que apresentam o termo “FUNCI” em seu nome juntas no Nível 1. Para o grupo 2, com comprovado conhecimento de domínio e do DW do estudo de caso, o resultado se aproximou um pouco mais do OntoDW. Entretanto, os especialistas do grupo 2 não realizaram a análise da relação entre os dados das colunas para a tarefa. Pelos questionários não terem fornecido resultados conclusivos, a regra R4 não pode ser considerada bem-sucedida. Entretanto, como o uso do conhecimento do domínio se mostrou importante, também se pode sugerir para a regra R4 um refinamento para que se utilize informações do metamodelo de domínio na sua definição. Para as regras R3, R8 e R10, conseguiu-se explicitar um conhecimento difundido entre analistas de BI no sentido de definir critérios para a realização de análises uniformes de modelos multidimensionais de dados.

5.3.4 Resultados das regras R7 e R9

A quarta seção de perguntas foi definida para avaliar as regras R7 e R9 do OntoDW. As regras R7 e R9 buscam identificar as instâncias das classes **RollUp** e **Hierarchy**, respectivamente, mas analisando os níveis de diferentes dimensões.

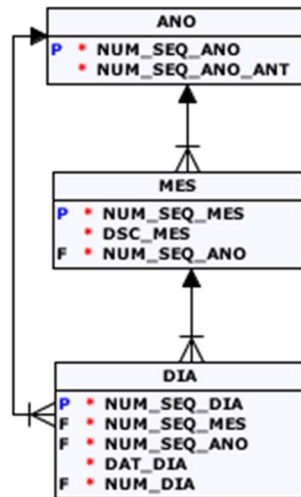


Figura 5.29: Recorte 3 do DW utilizado para validar as regras R7 e R9

06. Considerando que as tabelas apresentadas no Recorte 3 são dimensões, informe o número de operações possíveis de Roll Up que podem ser realizadas entre elas: *

Sua resposta

07. Caso tenha encontrado Roll Up's na tabela do Recorte 3, eles poderiam ser agrupados em uma hierarquia?: *

- Sim
- Não

Figura 5.30: Questões para validar as regras R7 e R9

As informações presentes no Recorte 3 foram suficientes para responder às questões acima? De qual informação sentiu falta?

Sua resposta

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

Sua resposta

Figura 5.31: Questões subjetivas sobre a seção 6 do formulário para o grupo 2

O questionário incluiu o recorte do modelo do DW ilustrado na Figura 5.29 e as questões apresentadas na Figura 5.30. Além disso, para colher os comentários do grupo 2 a respeito desta seção do formulário, foram feitas as questões subjetivas apresentadas na Figura 5.31. Para análise das respostas, elas foram agrupadas para comparação com o resultado obtido pela execução do OntoDW.

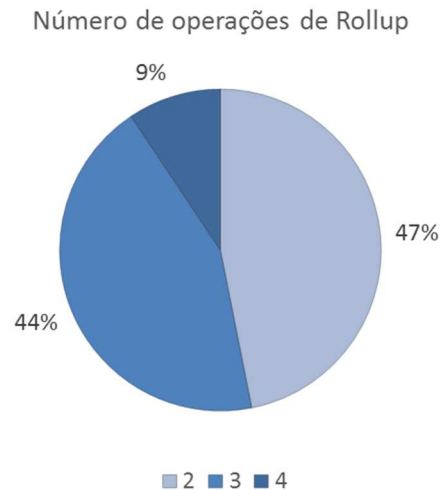


Figura 5.32: Respostas do grupo 1 para a questão 06

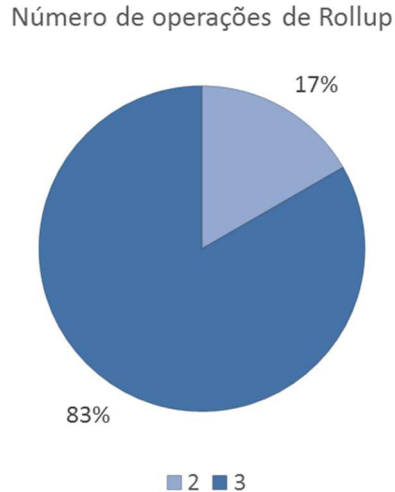


Figura 5.33: Respostas do grupo 2 para a questão 06

A questão 06 informa que as tabelas do Recorte 3 são tabelas de dimensão e solicita que seja informado o número de operações de *roll up* possíveis entre elas. Essa questão objetiva analisar o resultado da regra R7. Na Figura 5.32 e na Figura 5.33 são apresentadas as consolidações das respostas do grupo 1 e do grupo 2 para a questão 06, respectivamente. A execução do OntoDW encontrou 3 operações de *roll up* possíveis.

Analisando o resultado obtido com a questão 06, não houve uma quantidade de operações com maioria de escolha para o grupo 1, com as respostas se concentrando praticamente nos valores 2 e 3. Três (3) pessoas (9% do total) informaram encontrar 4 operações de *roll up* possíveis. A justificativa para um grande número de pessoas (47% do total) informar que encontraram 2 operações pode estar nos comentários fornecidos pelo grupo 2. Esse grupo apresentou respostas mais aderentes ao OntoDW e apenas 1 pessoa (17% do total) respondeu diferente, informando ter encontrado 2 operações.

Nos comentários foi citado pelas pessoas que houve hesitação em apontar um *roll up* entre as tabelas DIA e ANO, pois a semântica nos nomes das tabelas deixa muito clara a relação entre elas (tabelas de tempo) e esse *roll up* poderia não ser necessário pela relação através da tabela MES. Foi também comentado sobre a facilidade de responder esta questão pela relação ser explícita entre as tabelas e os conceitos serem de fácil compreensão.

O resultado apresentado pela regra R7 do OntoDW se mostra coerente por representar as possíveis operações de *roll up* que de fato estão implementadas no DW. O grupo 2 apresentou maioria de respostas para o mesmo valor do OntoDW (83% do total) e o conhecimento dos participantes sobre a implementação dos sistemas da organização sugere que auxiliou nas respostas.

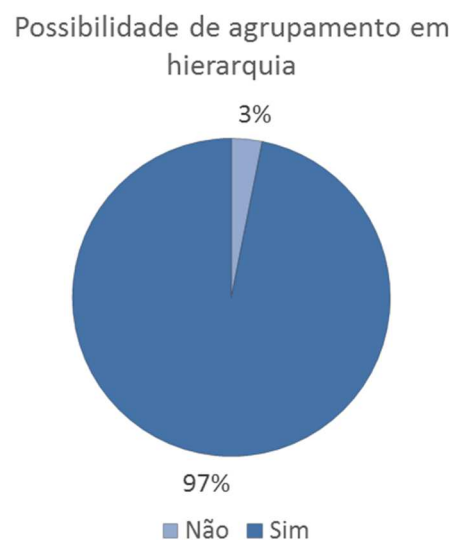


Figura 5.34: Respostas do grupo 1 para a questão 07

Possibilidade de agrupamento em hierarquia

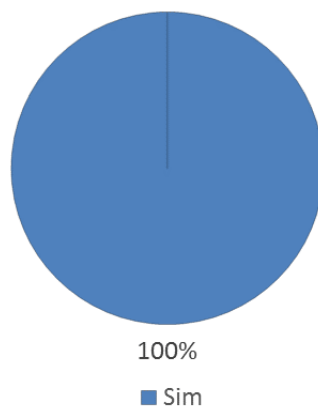


Figura 5.35: Respostas do grupo 2 para a questão 07

A questão 07 solicita que seja informado se os *roll ups* encontrados na Figura 5.29 podem ser agrupados em uma hierarquia. Essa questão objetiva analisar o resultado da regra R9. Na Figura 5.34 e na Figura 5.35 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2 para a questão 07, respectivamente.

É possível observar que o resultado para os dois grupos foi muito próximo, com a maioria das pessoas informando que é possível agrupar os *roll ups* encontrados em uma hierarquia. No grupo 1 foi atingido o índice de 97% de respostas “Sim” e no grupo 2 o índice de 100%. Pelos comentários do grupo 2 é possível concluir que a fácil compreensão dos conceitos representados pelas tabelas, e da relação entre elas, contribuiu pelo alto índice de respostas em comum. Sobre a única pessoa que marcou a opção “Não” no grupo 1, a única informação que talvez justifique sua escolha seja a baixa experiência em implementação de modelos multidimensionais (apenas 1 ano). O OntoDW mapeou uma hierarquia para os *roll ups* entre essas tabelas.

5.3.5 Resultados das regras R5, R6 e R11

A quinta seção de perguntas foi definida para avaliar as regras R5, R6 e R11 do OntoDW. As regras R5 e R6 buscam identificar as instâncias da classe **Measure** e a regra R11 busca identificar as instâncias da classe **SummarizabilityAlongFact**. A regra R6 é específica para tabelas de fato qualificadas como tabela de fato sem fato.

| FUNCI_3 | FUNCI | REG_ULT_MES_CARGA |
|------------------------------|------------------------------|-------------------------|
| PF * NUM_SEQ_ANO | PF * NUM_SEQ_MES | P * SIG_TEMA |
| PF * NUM_MATRIC_FUNCII | PF * NUM_MATRIC_FUNCII | P * NUM_SEQ_TIPO_PARAM |
| PF * COD_IDADE | F * COD_IDADE | * DSC_PARAM |
| PF * NUM_SEQ_ESTADO_CIVIL | F * NUM_SEQ_ESTADO_CIVIL | F NUM_SEQ_ULT_MES_CARGA |
| PF * NUM_SEQ_MUNIC | F * NUM_SEQ_MUNIC | |
| PF * NUM_SEQ_SITUAC_FUNCII | F * NUM_SEQ_PLANO_BENEF | |
| PF * NUM_SEQ_PLANO_BENEF | F * NUM_SEQ_PERIOD_NORMAT | |
| PF * NUM_SEQ_PERIOD_NORMAT | F * NUM_SEQ_CAPAC_CIVIL | |
| PF * NUM_SEQ_CAPAC_CIVIL | * QTD_FUNCII | |
| P * IND_RECBTO_APOSE | * QTD_DEP | |
| * QTD_FUNCII | * QTD_PENSTA | |
| * QTD_BENEF | * QTD_DESFIL | |
| * QTD_PENSTA | * VAL_SALAR_PARTIC | |
| * VAL_SALAR_PARTIC | * NUM_DIA_CONTRI_INSS_FORA | |
| * NUM_DIA_CONTRI_INSS_PATROC | * NUM_DIA_CONTRI_INSS_PATROC | |
| | * IND_RECBTO_APOSE | |

Figura 5.36: Recorte 4 do DW utilizado para validar as regras R5, R6 e R11

08. Informe a quantidade de medidas presentes na tabela
FUNCI_3: *

Sua resposta _____

09. Informe a quantidade de medidas presentes na tabela
FUNCI: *

Sua resposta _____

10. Informe a quantidade de medidas presentes na tabela
REG_ULT_MES_CARGA: *

Sua resposta _____

11. Caso tenha encontrado medidas nas tabelas do Recorte 4,
informe quantas dessas medidas estão presentes em mais de
uma tabela: *

Sua resposta _____

Figura 5.37: Questões para validar as regras R5, R6 e R11

As informações presentes no Recorte 4 foram suficientes para
responder às questões acima? De qual informação sentiu falta?

Sua resposta _____

Comente sobre a dificuldade de responder às questões acima, a
estratégia para chegar às respostas ou qualquer outro assunto
que achar pertinente.

Sua resposta _____

Figura 5.38: Questões subjetivas sobre a seção 7 do formulário para o grupo 2

Esta seção incluiu o recorte do modelo do DW ilustrado na Figura 5.36 e as questões apresentadas na Figura 5.37. Além disso, para colher os comentários do grupo 2 a respeito desta seção do formulário, foram feitas as questões subjetivas apresentadas na Figura 5.38. Para análise das respostas, elas foram agrupadas para comparação com o resultado obtido pela execução do OntoDW.

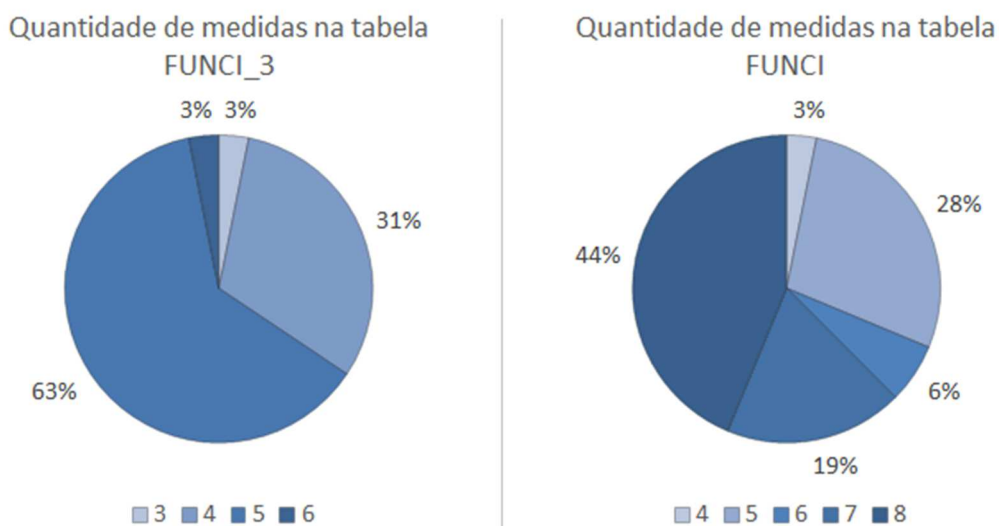


Figura 5.39: Respostas do grupo 1 para as questões 08 e 09

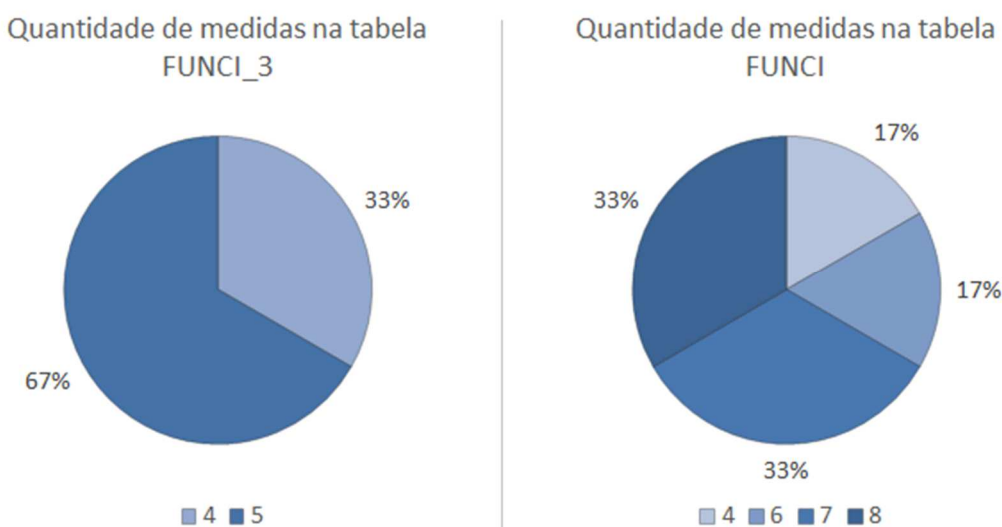


Figura 5.40: Respostas do grupo 2 para as questões 08 e 09

A questões 08 e 09 solicitam que seja informada a quantidade de medidas nas tabelas FUNCI_3 e FUNCI, respectivamente, e ambas visam analisar o resultado da regra R5. Na Figura 5.39 e na Figura 5.40 são apresentadas a consolidação das respostas

do grupo 1 e do grupo 2, respectivamente, para estas questões. A execução do OntoDW encontrou 5 medidas na tabela FUNCI_3 e 8 medidas na tabela FUNCI.

Para a tabela FUNCI_3, as respostas dos dois grupos foram semelhantes, proporcionalmente. O valor de 5 medidas como opção da maioria dos respondentes (em torno de 2/3 para ambos os grupos) e o valor de 4 medidas praticamente para os respondentes restantes (em torno de 1/3 para ambos os grupos). Uma das justificativas possíveis para o número considerável de respostas para 4 medidas é o nome da coluna NUM_DIA_CONTRI_INSS_PATROC. Na comparação com as outras colunas da tabela, os termos utilizados em seu nome sugerem que seja uma chave estrangeira pela sua similaridade com essas colunas. Entretanto, a coluna armazena o número de dias de contribuição do funcionário para o INSS, estando vinculado ao patrocinador do fundo de pensão. Os comentários registrados pelo grupo 2 reforçam essa possibilidade, ao encontrarmos referências explícitas a essa coluna e também sobre a dúvida pela ausência da sinalização de chave estrangeira. A resposta escolhida pela maioria foi a mesma obtida pelo OntoDW.

Para a tabela FUNCI, as respostas dos dois grupos também foram semelhantes entre si, mas divergentes do padrão encontrado para a tabela FUNCI_3: as respostas foram mais variadas e nenhuma das opções foi escolhida pela maioria de nenhum dos dois grupos. Entretanto, a resposta mais frequente nos 2 grupos foi a de que existem 8 medidas na tabela, a mesma obtida pelo OntoDW. Além da justificativa descrita no parágrafo anterior sobre a coluna NUM_DIA_CONTRI_INSS_PATROC ser pertinente também para esta tabela, temos a coluna NUM_DIA_CONTRI_INSS_FORA com a mesma característica e a coluna IND_RECBTO_APOSE que não deixa claro em seu nome o tipo de dado que armazena, o que pode auxiliar para definir a coluna como medida.

A coluna IND_RECBTO_APOSE apresentou outra situação nessa tabela que pode ter dificultado a tarefa de análise. Ela está presente na tabela FUNCI_3 como parte da chave primária e na tabela FUNCI fora da chave primária. Dessa forma, cada especialista poderia priorizar um aspecto e definir a coluna como medida ou não. Essa questão também foi registrada pelo grupo 2 em seus comentários. Para esclarecer a questão, foi consultado um responsável pelo sistema e foi constatado que se trata de um erro de implementação. Nas duas tabelas essa coluna deve fazer parte da chave primária. Se forem consideradas as opções entre 5 e 8 medidas (por existirem situações de possíveis dúvi-

das dos respondentes para 3 colunas), atingimos 97% das respostas para o grupo 1 e 83% das respostas para o grupo 2.

A análise das questões indicou que o resultado gerado pela execução da regra R5 é coerente e que a tarefa de identificação das classes nas tabelas é bem subjetiva, reforçando a importância da definição de regras para prover uma padronização da interpretação de modelos.

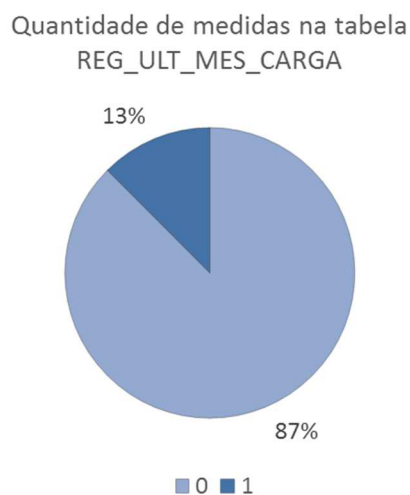


Figura 5.41: Respostas do grupo 1 para a questão 10

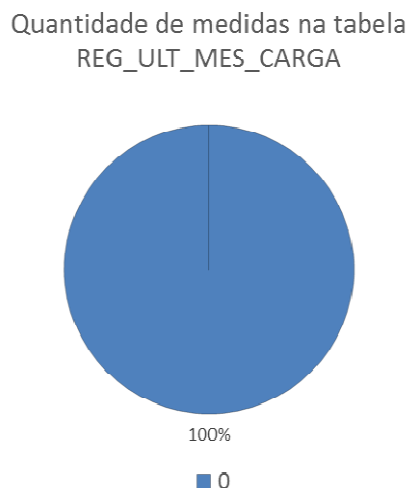


Figura 5.42: Respostas do grupo 2 para a questão 10

A questão 10 solicita que seja informada a quantidade de medidas presentes na tabela REG_ULT_MES_CARGA, visando avaliar o resultado da regra R6. Na Figura 5.41 e na Figura 5.42 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2, respectivamente, para esta questão.

A execução do OntoDW mapeou 1 medida a partir da tabela REG_ULT_MES_CARGA, mas as respostas dos dois grupos foram muito diferentes disso. Apenas 4 pessoas do grupo 1 (13% do total) informou ter encontrado 1 medida e no grupo 2 nenhum dos respondentes encontrou alguma. A justificativa mais clara para isso é o fato de que esta tabela foi erroneamente localizada no DW, já que a mesma é uma tabela de controle, e não uma tabela de fato. Para as pessoas que não consideraram REG_ULT_MES_CARGA como uma tabela de fato, não faz sentido que a mesma contenha medidas. Uma outra justificativa possível é a falta de colunas com as características comumente encontradas em medidas, como a ausência de chaves, a indicação de um tipo de dados sumarizável (numeral) e termos de negócio no nome. Dessa forma, não fica explícita a existência de uma medida. Na Seção 5.2 foi descrito o caso desta tabela e concluído que o resultado da regra R6 é pertinente.

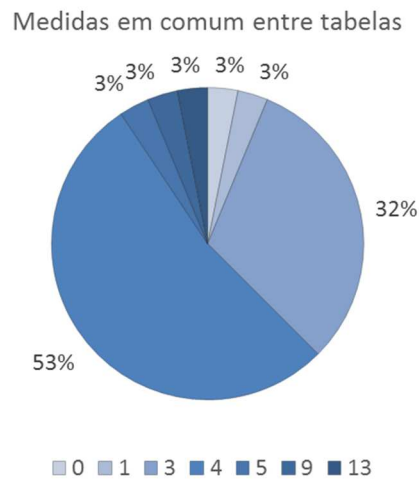


Figura 5.43: Respostas do grupo 1 para a questão 11

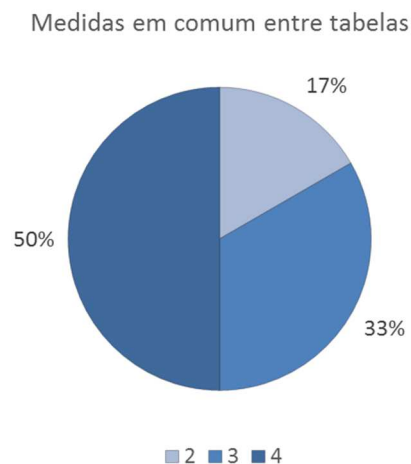


Figura 5.44: Respostas do grupo 2 para a questão 11

A questão 11 solicita que seja informado a quantidade de medidas presentes em mais de uma tabela da Figura 5.36, visando avaliar o resultado da regra R11. Na Figura 5.43 e na Figura 5.44 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2, respectivamente, para esta questão.

A execução do OntoDW mapeou 4 medidas em comum entre as tabelas FUNC1_3 e FUNC1. As respostas dos dois grupos foram semelhantes entre si e a mesma resposta do OntoDW foi escolhida pela maioria, mas que representou apenas metade dos respondentes dos grupos. Em cada grupo, também foi informado o número de 3 medidas por aproximadamente 1/3 das pessoas e outras opções pelo restante dos respondentes (aproximadamente 1/6). Para esta questão foram encontradas também algumas respostas incoerentes, sugerindo que o respondente não entendeu o que se pedia. Um exemplo foi a resposta de 13 métricas em comum, quando esta pessoa informou ter encontrado no máximo 8 medidas em uma das tabelas.

A justificativa encontrada para o número considerável de respostas para 3 medidas em comum é novamente a nomenclatura da coluna NUM_DIA_CONTRI_INSS_PATROC, que fez aproximadamente 1/3 dos respondentes dos dois grupos não considerá-la como medida.

Após esta análise, a regra R11 foi considerada bem-sucedida, mas novamente foi observada a facilidade de se obter resultados diferentes por conta da alta subjetividade da tarefa ao ser realizada por analistas.

5.3.6 Resultados das regras R12 e R13

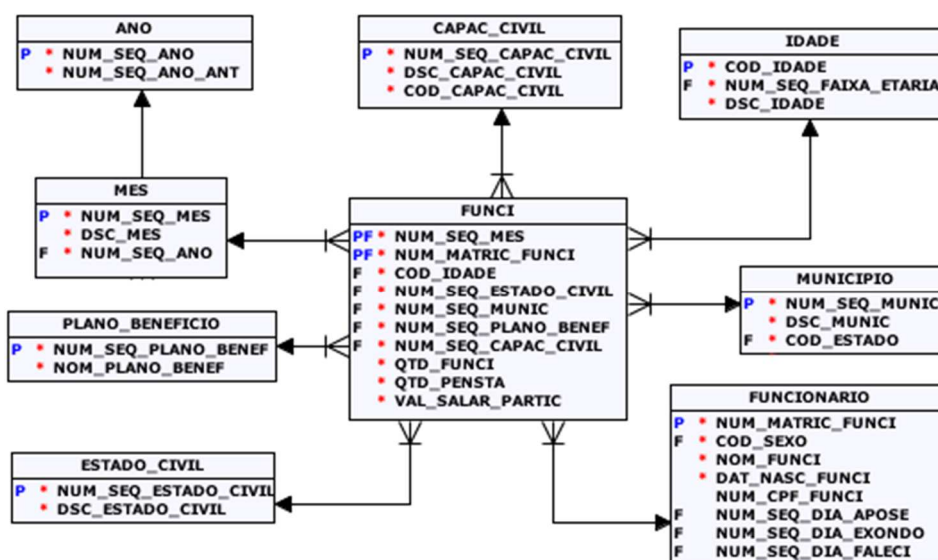


Figura 5.45: Recorte 5 do DW utilizado para validar as regras R12 e R13

A sexta e última seção de perguntas foi definida para avaliar as regras R12 e R13 do OntoDW. A regra R12 busca identificar as instâncias da classe **SummarizabilityAlongDimension** e a regra R13 busca identificar as instâncias da classe **SummarizabilityAlongHierarchy**.

12. Caso encontre medidas nas tabelas do Recorte 5, informe a quantidade de dimensões distintas que podem ser usadas para analisar essas medidas: *

Sua resposta

13. Caso encontre medidas nas tabelas do Recorte 5, informe a quantidade de hierarquias distintas que podem ser usadas para analisar essas medidas: *

Sua resposta

Figura 5.46: Questões para validar as regras R12 e R13

As informações presentes no Recorte 5 foram suficientes para responder às questões acima? De qual informação sentiu falta?

Sua resposta

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

Sua resposta

Figura 5.47: Questões subjetivas sobre a seção 8 do formulário para o grupo 2

Esta seção foi incluída o recorte do modelo do DW ilustrado na Figura 5.45 e as questões apresentadas na Figura 5.46. Para colher os comentários do grupo 2 a respeito desta seção do formulário, foram também feitas as questões subjetivas apresentadas na Figura 5.47. Para análise das respostas, elas foram agrupadas para comparação com o resultado obtido pela execução do OntoDW.

A questão 12 solicita que, a partir do recorte do DW da Figura 5.45, seja informada a quantidade de dimensões distintas que podem ser utilizadas para analisar as medidas presentes no modelo. Essa questão objetiva avaliar o resultado da regra R12. Na Figura 5.48 e na Figura 5.49 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2, respectivamente, para esta questão.

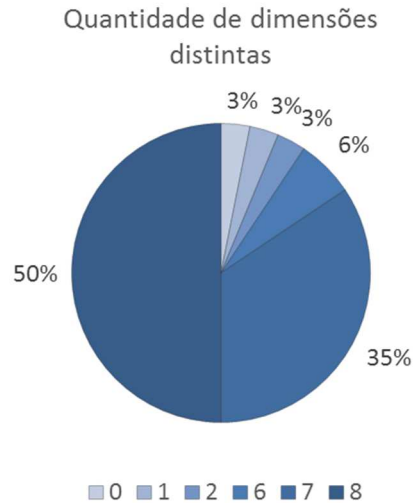


Figura 5.48: Respostas do grupo 1 para a questão 12

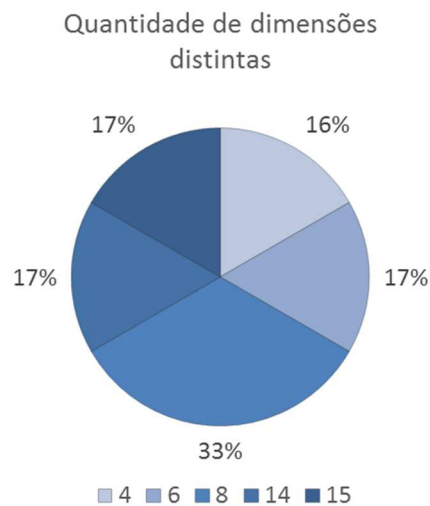


Figura 5.49: Respostas do grupo 2 para a questão 12

A execução do OntoDW mapeou 8 dimensões distintas para estas tabelas apresentadas e as respostas dos dois grupos para esta questão foram bastante distintas entre si. Para o grupo 1, metade das pessoas informou identificar 8 dimensões (mesma resposta do OntoDW) e 35% informou a quantidade de 7 dimensões. Uma justificativa encontrada para o número considerável de respostas para 7 dimensões foi que a tabela ANO não se relaciona diretamente com a tabela de fato, através de chave estrangeira. Por este motivo, algumas pessoas podem considerá-la apenas como um agrupamento de MES, e não como uma análise possível das medidas.

Para o grupo 2, foi encontrada uma distribuição mais uniforme das respostas pelas opções respondidas. A resposta do OntoDW também foi a mais escolhida para este

grupo, mas representou somente 33% das respostas. Ao analisar os comentários fornecidos pelo grupo, foram extraídas algumas informações que auxiliam a entender as respostas de uma parte do grupo. Duas (2) pessoas informaram um número de dimensões maior que o número de tabelas presente no recorte do DW, mas isso porque elas consideraram as chaves estrangeiras sinalizadas em algumas colunas das dimensões. Assim, contaram como se as tabelas referenciadas estivessem no modelo. As 2 pessoas restantes informaram um número de dimensões para o qual não foi encontrado o raciocínio utilizado. No entanto, é relevante citar que essas são as duas pessoas menos experientes deste grupo e a questão 12 apresenta a tarefa do formulário de pesquisa que envolve mais classes, o que pode torná-la confusa para algumas pessoas. Isso porque, para responder à questão, o respondente deve identificar no modelo os fatos e dimensões, analisar se existem medidas e quais dimensões podem ser utilizadas na análise das medidas. É possível que essas 2 pessoas não tenham compreendido corretamente a tarefa solicitada ou tenham se confundido ao analisar o modelo para responder a questão.

Ao analisar as respostas dos grupos, não foi encontrada alguma situação que tornasse a resposta do OntoDW incoerente. Assim, mesmo não conseguindo obter com as pesquisas uma grande concentração das respostas esperadas, pode-se considerar a regra R12 bem-sucedida.

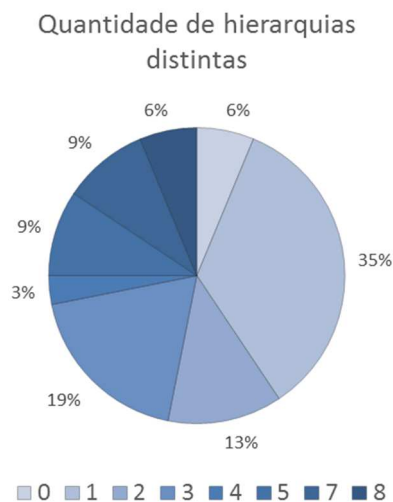


Figura 5.50: Respostas do grupo 1 para a questão 13

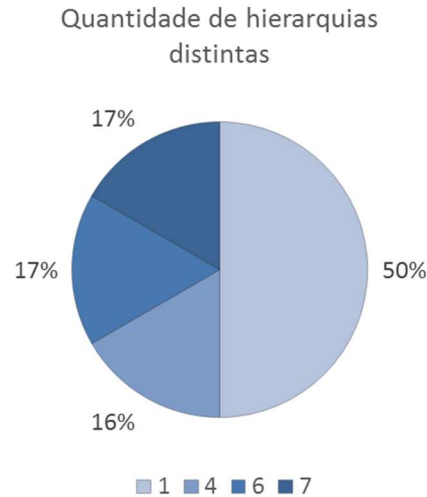


Figura 5.51: Respostas do grupo 2 para a questão 13

A questão 13 solicita que seja informada a quantidade de hierarquias distintas podem ser utilizadas para analisar as medidas presentes no mesmo modelo da questão anterior. Essa questão objetiva avaliar o resultado da regra R13. Na Figura 5.50 e na Figura 5.51 são apresentadas a consolidação das respostas do grupo 1 e do grupo 2, respectivamente, para esta questão.

A execução do OntoDW mapeou 2 hierarquias distintas para estas tabelas apresentadas, uma contendo os níveis das tabelas MES e ANO e outra hierarquia com os níveis da tabela FUNCIONARIO. As respostas dos dois grupos para esta questão também foram bastante distintas entre si.

Para o grupo 1, o fato inesperado foi que 35% das pessoas (nesse caso também a opção mais escolhida pelo grupo) informou não ter encontrado nenhuma hierarquia no modelo. Isto foi surpreendente pois uma das hierarquias era explícita ao utilizar chaves estrangeiras e porque uma estrutura semelhante havia sido apresentada em uma seção anterior do formulário e as respostas obtidas foram coerentes. O restante das respostas foi dividido por valores entre 1 e 8 sem concentração relevante em nenhuma opção específica. Isso pode ser explicado pela dependência da resposta da questão anterior, onde o respondente identificou as dimensões do modelo que poderiam compor as hierarquias, e as respostas também foram esparsadas, e pelas diferentes possibilidades de análise que podem ter seguido cada analista, como contabilizar ou não as chaves estrangeiras que referenciam tabelas não apresentadas ou os níveis de dimensão identificados.

Para o grupo 2, as respostas também foram coerentes com as respostas da questão anterior. As pessoas que haviam informado um grande número de dimensões encon-

tradas também informaram um alto número de hierarquias. Além delas, metade do grupo apresentou a mesma resposta do OntoDW e 1 pessoa restante, no caso a menos experiente do grupo, informou um número de hierarquias (4) para o qual não foi possível encontrar o raciocínio utilizado. Diferente do grupo 1, ninguém informou não ter encontrado hierarquia. Possivelmente o conhecimento do ambiente tecnológico da organização auxiliou, já que as pessoas já conhecem a forma que normalmente as hierarquias são implementadas.

Ao analisar as respostas dos grupos, também não foi encontrada para a regra R13 questões que tornassem a resposta do OntoDW incoerente. Assim, a regra R12 foi também considerada bem-sucedida. Dessa forma, também se conseguiu explicitar para estas regras um conhecimento difundido entre analistas de BI no sentido de definir critérios para realização de análises uniformes de modelos multidimensionais de dados.

5.3.7 Considerações sobre os resultados

Este capítulo apresentou os resultados consolidados da aplicação dos dois formulários criados para a validação das regras de mapeamento propostas no presente trabalho. Foram também apresentadas as respectivas avaliações dessas respostas, juntamente com a comparação entre os grupos de respondentes e com o resultado gerado pelo OntoDW.

Foi possível observar o alto teor de subjetividade da tarefa de análise dos modelos de dados, fazendo com que fossem obtidos resultados muito diversos. Nas questões envolvendo conceitos mais difundidos e de visualização mais explícita, como fatos e dimensões, os resultados foram mais homogêneos. Nas questões envolvendo a identificação de conceitos menos explícitos ou com aplicação mais restrita em esquemas de dados, como hierarquias ou fatos sem fato, ocorreu uma maior variedade de respostas. Até a organização das tabelas e colunas no modelo foi citada como importante para a realização da análise. Isso mais uma vez reforça a necessidade de padronização dos conceitos e a adoção de um metamodelo OLAP compartilhado, o que permitiria uma maior convergência nos resultados e aumento do potencial de análise do ambiente de BI.

Em muitos casos, uma não convergência com o OntoDW não representou necessariamente que existia algum erro ou problema. Um modelo multidimensional é criado com o intuito de organizar indicadores por assunto para apoiar tomada de decisão de

forma ágil. Técnicas que seriam não recomendadas em sistemas transacionais são comumente encontradas em *Data Warehouses*, como duplicação de informação. Assim, um especialista pode ter a experiência de implementar uma coluna em uma tabela fato sem chave estrangeira para representar uma dimensão de muito baixa cardinalidade (como Sexo) o que outro analista identificaria como uma medida, pela sua estrutura. Dependendo do contexto da aplicação, ambos os raciocínios estariam corretos e, portanto, faz-se necessário estabelecer semânticas precisas e diretrizes de implementação de esquemas conceituais mais bem definidas.

É importante ressaltar também a complexidade e tamanho dos questionários aplicados. Além das questões para definição de perfil, os questionários incluem 5 diferentes recortes de modelo para a identificações de 13 conceitos relativos a modelagem multidimensional, com questões para associação de instâncias, contagem de resultados e de cruzamento de informações. Os recortes incluídos também apresentam diferentes técnicas empregadas na modelagem de *Data Warehouses*, que podem não ser de domínio por todos os respondentes na mesma escala.

Conforme descrito até então, foi concluído que 11 dentre as 13 regras de mapeamento propostas foram bem-sucedidas ao identificar os conceitos na aplicação deste estudo de caso. No entanto, os comentários colhidos, juntamente com as análises realizadas, indicaram os termos extraídos dos metadados das tabelas podem ser utilizados não somente para a definição de nomes de instâncias, mas também para a identificação de instâncias e relações da ontologia de aplicação gerada.

A semântica presente no nome dos objetos foi comentada como muito importante pelos analistas para responder as questões, tanto para identificar conceitos de negócio e suas relações (mês e ano, por exemplo) quanto para identificar a função que o objeto está desempenhando (fato e medida, por exemplo).

5.4 Análise da utilidade da ontologia gerada por usuários

Para validar a premissa de que uma representação do conhecimento pode apoiar a análise dos dados do DW, foi aplicado um questionário online com usuários do sistema objeto do estudo de caso para avaliar a utilidade da ontologia gerada pelo OntoDW sob o ponto de vista do usuário. Esse questionário online também foi aplicado com a utilização do google forms e utilizando a escala Likert para as questões objetivas relaci-

onadas ao perfil do respondente. No entanto, este questionário contou com maioria de questões subjetivas, para colher opiniões dos respondentes. O questionário completo aplicado se encontra no ANEXO III.

A seguir é detalhada a aplicação do questionário, incluindo o detalhamento da reunião prévia realizada com o grupo, o número e o perfil dos respondentes, as questões apresentadas e a análise dos resultados. A análise dos resultados é qualitativa e busca identificar os pontos em destaque relatados sobre os trechos apresentados da ontologia gerada pelo OntoDW. A totalidade das respostas subjetivas deste questionário se encontra no ANEXO V.

5.4.1 Reunião prévia com os usuários

Antes do envio deste questionário, foi realizada uma reunião com o grupo de participantes para apresentar mais detalhes sobre o trabalho desenvolvido e o formulário que seria respondido por eles e dirimir eventuais dúvidas. Assim sendo, foi apresentada uma versão impressa do questionário e explicado que seu objetivo era obter opiniões sobre a utilidade dos recortes da ontologia para apoiar a tarefa de analisar os dados do DW. Os recortes da ontologia foram representados no questionário graficamente e através de capturas de tela no Protégé. A notação adotada pela ferramenta foi explicada com o auxílio do questionário impresso. Para auxiliar a explicação da tela do Protégé, foi utilizado um notebook para demonstração do seu funcionamento. A ontologia avaliada foi a mesma gerada pelo OntoDW no estudo de caso.

A demonstração possibilitou que o grupo observasse uma navegação pelos conceitos da ontologia através do Protégé. Nessa ocasião, foi também possível registrar alguns comentários emitidos:

- A visualização gráfica da ontologia no Protégé é confusa devido à quantidade de objetos que aparecem ao realizar a navegação pelo OntoGraf, sem nenhuma organização;
- O Protégé fica excessivamente lento ao incluir objetos na representação do Ontograf;
- Existem muito poucas funcionalidades para organizar o modelo;
- A tela de apresentação das instâncias agradou mais que a representação gráfica. Os principais motivos foram a possibilidade de selecionar a classe como filtro das instâncias apresentadas e a possibilidade de utilizar as relações entre as ins-

tâncias como links entre elas. Utilizando esses links foi possível navegar por toda a ontologia de forma rápida. O ponto negativo é que as relações são apresentadas sem organização ou classificação;

- Este tipo de documentação do sistema tem potencial para ser aplicado em iniciativas de gestão do conhecimento na organização, pois deixa mais claro que informações estão disponíveis;
- Seria útil se a ferramenta estivesse disponível de imediato para utilização;

5.4.2 Perfil

A primeira seção de perguntas neste questionário foi definida com o objetivo de analisar o perfil dos respondentes em relação ao conhecimento e experiência com aplicações de BI e obter mais informações sobre a relação dos mesmos com a organização.

Qual seu domínio nos conceitos de aplicações de BI (ex.: hierarquia, atributo, métrica)?

(3 respostas)

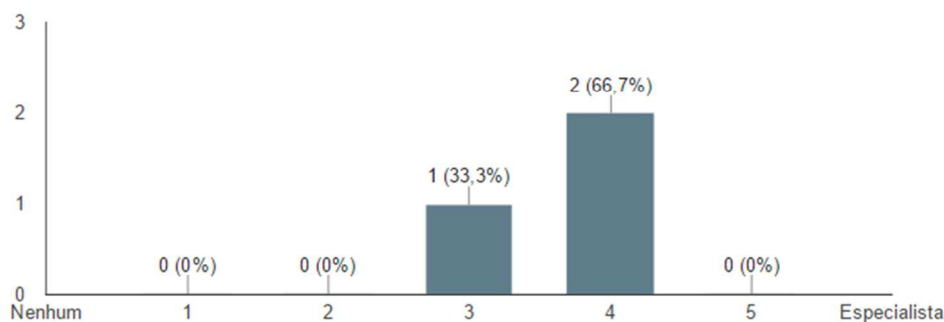


Figura 5.52: Perfil do grupo 3 em relação ao domínio em conceitos de aplicações de BI

Esse questionário foi destinado a usuários de sistema de BI, e foi respondido por 3 pessoas. Este grupo de respondentes será chamado neste trabalho de grupo 3. A Figura 5.52 apresenta o grau de domínio dos respondentes do grupo 3 nos conceitos de aplicações de BI. É possível considerar que o grupo domina os conceitos pelo fato da maioria das pessoas pelo menos terem conhecimento intermediário. Vale ressaltar que os respondentes deste grupo são usuários do sistema objeto do estudo de caso e utilizam os dados do *Data Warehouse* através de uma ferramenta OLAP. Assim sendo, esse conhecimento é aplicado nos conceitos identificados pelo OntoDW, o que torna as respostas colhidas mais próximas do que seria de fato o uso da ontologia gerada no apoio ao uso do sistema de BI.

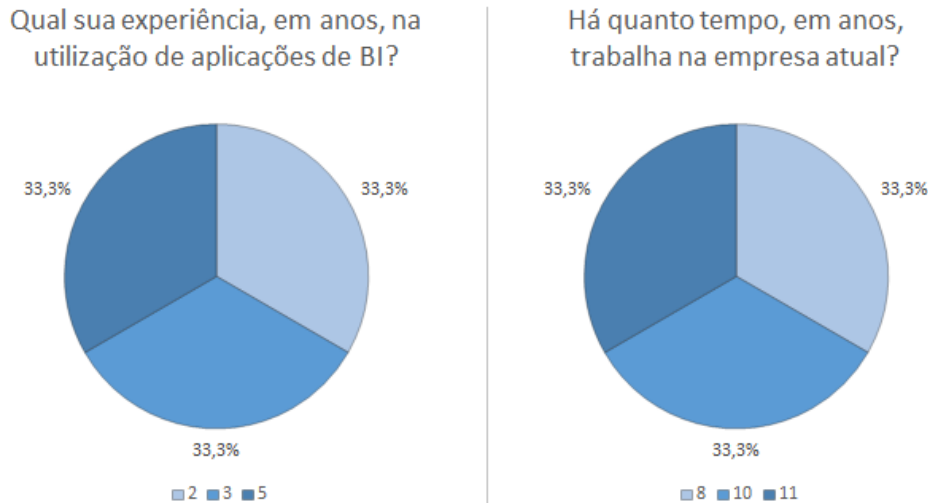


Figura 5.53: Perfil do grupo 3 em relação à sua experiência

A Figura 5.53 apresenta o tempo de experiência em anos dos respondentes do questionário 3 na utilização de aplicações de BI (gráfico à esquerda) e tempo de trabalho na organização atual (gráfico à direita). Os gráficos auxiliam a definir esse grupo de respondentes como experiente na utilização das aplicações de BI da organização, pois informaram utilizar aplicações de BI por no mínimo 2 anos e terem pelo menos 8 anos de trabalho, o que é um indício de bom conhecimento do negócio. Por fim, os respondentes informaram seus cargos: enquanto 1 deles é gerente de núcleo, os outros dois são analistas. Todos trabalham com atuária e gestão do cadastro dos funcionários.

5.4.3 Resultados

Após a primeira seção do formulário, são apresentadas questões subjetivas com o objetivo de colher os comentários dos usuários. A segunda seção contém questões relativas à opinião do respondente acerca da utilidade de representações de conhecimento de sistemas de BI de forma mais genérica. Outras 4 seções foram definidas com foco mais específico no estudo de caso. Em cada uma destas seções é apresentado um recorte da ontologia gerada pelo OntoDW e, a partir deste recorte, são apresentadas questões sobre a utilidade do recorte apresentado. Os conceitos de modelagem multidimensional ou de aplicações OLAP citados são definidos no topo de cada seção, de forma a unificar o entendimento dos usuários antes que respondam as perguntas.

Qual sua opinião sobre a utilidade de uma representação visual que descrevesse as informações implementadas no DW alinhadas com conceitos de BI? Ex.: Demonstrassem as dimensões e as medidas (métricas) disponíveis, as análises possíveis de se realizar. *

Sua resposta

Existiria vantagem entre uma representação gráfica e uma textual? *

Sua resposta

Se essas informações fossem apresentadas com o uso de termos próprios do negócio ao invés dos nomes no DW, seriam mais úteis? *

Sua resposta

Figura 5.54: Questões sobre a utilidade de representações de conhecimento

Na segunda seção, foi solicitado aos usuários que as questões fossem respondidas com base em sua experiência de utilização de aplicações de BI e nas necessidades e possibilidades de melhoria que identifique nas atividades de trabalho que envolvam análise e geração de informações. As 3 questões desta seção estão apresentadas na Figura 5.54.

A primeira destas questões solicitava uma opinião sobre a utilidade de uma representação visual que descrevesse as informações implementadas no DW alinhadas com conceitos de BI. Um dos respondentes citou diversas vantagens na utilização de uma representação visual, mas focou na representação dos dados do DW de forma visual. Assim, não foi possível extrair da sua resposta sua opinião sobre a representação de informações possíveis de serem utilizadas. Entretanto, os outros dois respondentes compreenderam melhor o objetivo da pergunta. Em suas respostas, informaram que essa representação visual atuaria como catalisador do aprendizado sobre o negócio que está modelado no BI e que seria de grande valia, pois permitiria ao usuário explorar melhor a ferramenta OLAP disponível.

A segunda questão buscava opiniões sobre preferências entre representações gráficas ou textuais. A opinião geral é de que uma representação gráfica seria mais vantagio-

sa. Foi comentado que ela seria de mais fácil compreensão que uma representação textual e teria a vantagem de sintetizar e facilitar a percepção das informações.

A terceira e última questão desta seção busca opiniões sobre a utilidade de termos de negócio na representação de informações. A opinião comum é que seria mais útil o uso de termos próprios do negócio. Foi comentado que, para o usuário, facilitaria o entendimento e tornaria mais intuitiva a aplicação dos dados contemplados no DW e que, quanto mais próximo o BI estiver da linguagem do negócio, mais à vontade e seguros atuarão os usuários finais.

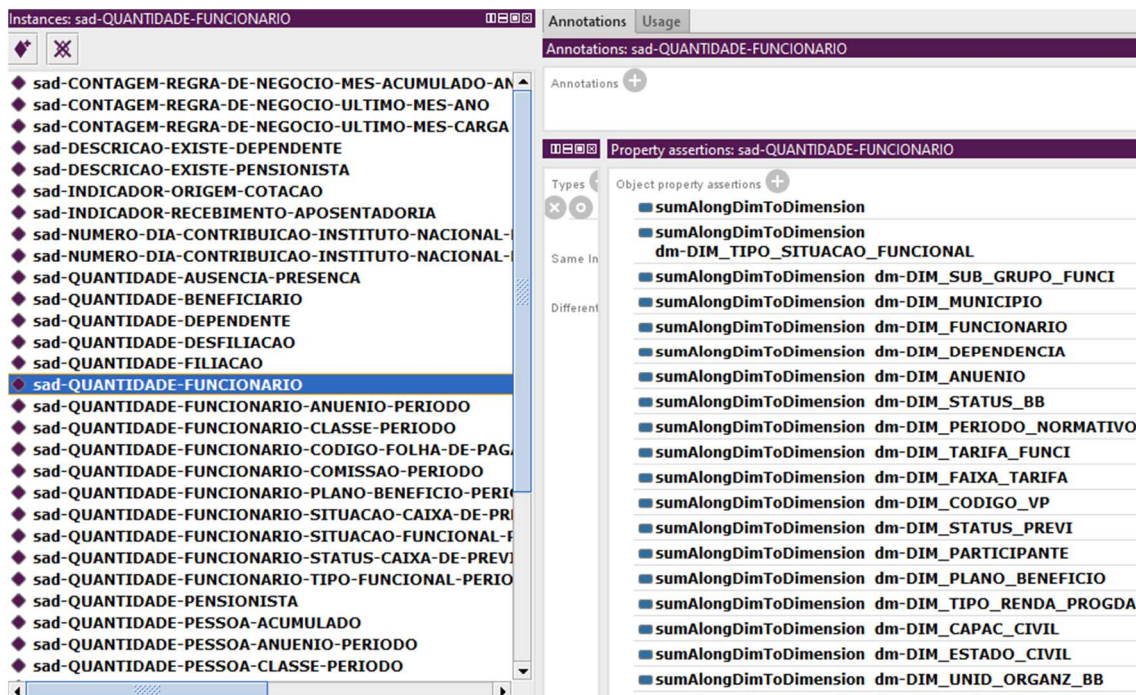


Figura 5.55: Recorte 1 da ontologia para a pesquisa com usuários

01. O Recorte 1 mostra as dimensões disponíveis para se analisar (lado direito da imagem) a medida/métrica QUANTIDADE-FUNCIONARIO (lado esquerdo da imagem). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW. *

Sua resposta

02. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 1, como a inclusão de informações ou sua organização?

Sua resposta

Figura 5.56: Questões sobre o recorte 1 da ontologia

Na terceira seção do formulário é apresentada uma captura de tela do Protégé com a ontologia gerada pelo OntoDW, onde estão sendo apresentadas dimensões para uso na análise da métrica de quantidade de funcionários selecionada. Essa captura de tela é ilustrada na Figura 5.55 e as questões formuladas para colher as opiniões do grupo 3 sobre essa captura estão representadas na Figura 5.56.

A questão 01 descreve o recorte 1 e solicita a opinião sobre sua utilidade para a análise de dados do DW. A opinião colhida das pessoas do grupo 3 foi de que esse tipo de visualização permite ao usuário identificar as relações com mais facilidade e evita a utilização de uma métrica de forma ineficaz e a construção de pesquisas cujos resultados sejam inconsistentes. Foi também citado que seria muito útil durante a fase de aprendizagem do negócio e do ambiente do BI.

A questão 02 busca obter críticas ou sugestões sobre a representação de conhecimento apresentada no recorte 1. Nessa questão, as respostas foram diversas. Um usuário informou ter achado a representação satisfatória, outro informou que precisaria navegar pelo Protégé para emitir uma opinião mais colaborativa e o terceiro sugeriu que as relações das tabelas poderiam ser representadas de forma mais intuitiva, como já existem em alguns aplicativos disponíveis no mercado. Essa sugestão está relacionada à interface do Protégé, que os usuários acharam confusa na reunião prévia realizada.

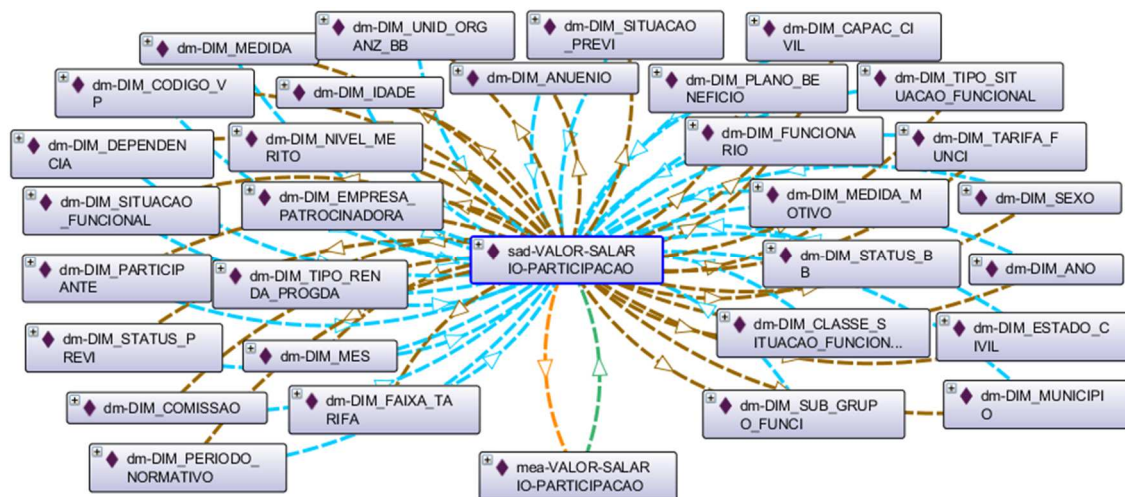


Figura 5.57: Recorte 2 da ontologia para a pesquisa com usuários

Na quarta seção do formulário é apresentado um recorte da ontologia gerada pelo OntoDW, ilustrado na Figura 5.57, onde são apresentadas as dimensões disponíveis para análise da métrica associada ao salário de participação do funcionário ou ex-

funcionário. As questões formuladas para colher as opiniões sobre esse recorte estão representadas na Figura 5.58.

03. O Recorte 2 mostra as dimensões disponíveis para se analisar (quadrados com nome iniciando com "dm-") a medida/métrica VALOR-SALARIO-PARTICIPACAO (quadrados com nome iniciando com "mea-"). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW. *

Sua resposta

04. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 2, como a inclusão de informações ou sua organização?

Sua resposta

Figura 5.58: Questões sobre o recorte 2 da ontologia

A questão 03 descreve o recorte 2 e solicita a opinião sobre sua utilidade para a análise de dados do DW. A opinião de dois usuários foi de que ele é útil, pois permite visualizar as combinações possíveis na construção das pesquisas e auxilia no entendimento das várias relações existentes entre os conceitos. O usuário restante opinou que a figura se apresentou muito confusa, comprometendo qualquer visualização lógica. Entretanto, esse usuário ressaltou que estava visualizando o recorte da ontologia através de um celular.

A questão 04 busca obter críticas ou sugestões sobre a representação de conhecimento apresentada no recorte 2. Para este recorte, foram feitas algumas sugestões pelos usuários, todas elas relativas à organização das instâncias na representação gráfica do Protégé:

- Estruturar os dados de acordo com algum critério a ser definido pelo usuário, pois existe uma sobreposição de informações na figura;
- Categorizar os dados criando classes mais abrangentes, o que traria ganho de entendimento;
- Incluir legenda para caracterizar as linhas tracejadas (relações).

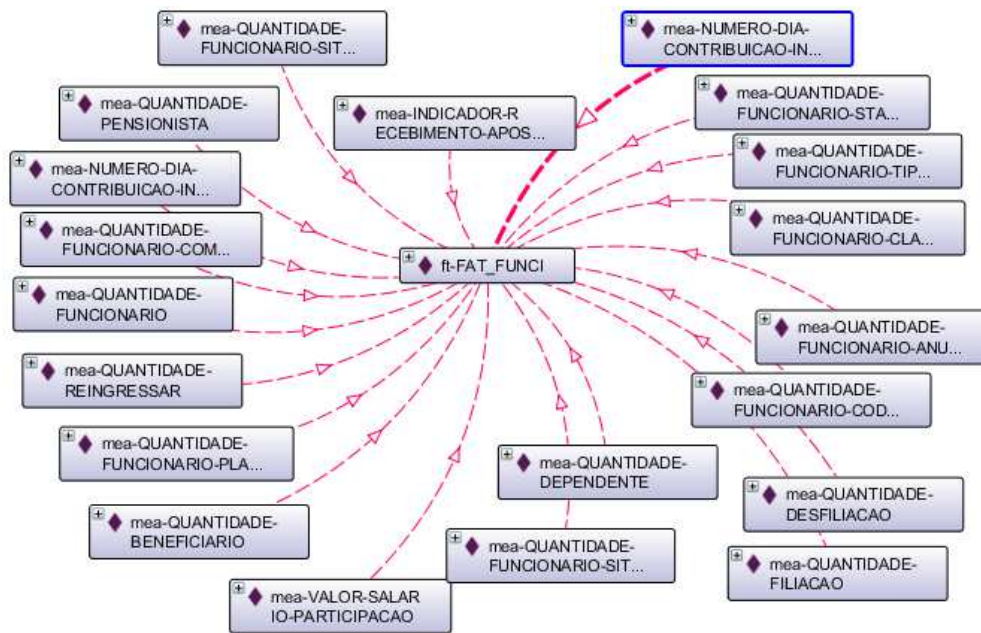


Figura 5.59: Recorte 3 da ontologia para a pesquisa com usuários

05. O Recorte 3 mostra as medidas/métricas (quadrados com nome iniciando com "mea-") contidas no fato FAT_FUNCI (quadrados com nome iniciando com "ft-"). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW. *

Sua resposta

06. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 3, como a inclusão de informações ou sua organização?

Sua resposta

Figura 5.60: Questões sobre o recorte 3 da ontologia

Na quinta seção do formulário é apresentado um recorte da ontologia gerada pelo OntoDW, ilustrado na Figura 5.59, onde são apresentadas as métricas contidas na tabela de fato FAT_FUNCI (representada na ontologia pela instância ft-FAT_FUNCI). As questões formuladas para colher as opiniões sobre esse recorte estão representadas na Figura 5.60.

A questão 05 descreve o recorte 3 e solicita a opinião sobre sua utilidade para a análise de dados do DW. As opiniões para este recorte foram diversas. Um usuário acredita que ele facilita o entendimento das inter-relações das medidas/métricas e, con-

sequentemente, a análise dos resultados obtidos. Outra resposta novamente falou da questão visual, opinando que se fosse utilizada uma ordenação alfabética no recorte facilitaria a identificação dos itens por usuários mais experientes. O terceiro respondente informou que este recorte é importante pois auxiliaria no entendimento da estrutura do sistema, o que está por traz da interface da ferramenta OLAP. Esta opinião está aderente com a descrição da proposta de solução na Seção 3, onde é citado que o resultado também é útil aos analistas de BI.

A questão 06 busca obter críticas ou sugestões sobre a representação de conhecimento apresentada no recorte 3. Para este recorte, as sugestões foram que existisse possibilidade do usuário estruturar as informações de acordo com suas necessidades e também ordenar as instâncias, como ordem alfabética por exemplo.

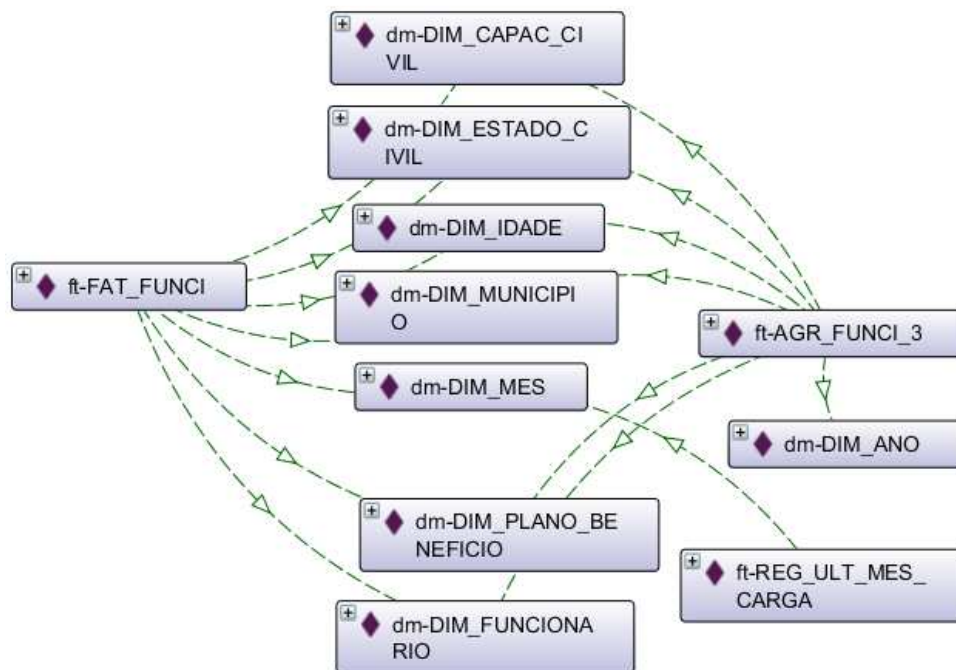


Figura 5.61: Recorte 4 da ontologia para a pesquisa com usuários

Na sexta e última seção de perguntas do formulário é apresentado outro recorte da ontologia gerada pelo OntoDW, ilustrado na Figura 5.61, onde são apresentadas dimensões e fatos relacionados. As questões formuladas para colher as opiniões sobre esse recorte estão representadas na Figura 5.62.

07. O Recorte 4 mostra as dimensões disponíveis (quadrados com nome iniciando com "dm-") inter-relacionadas com os fatos disponíveis (quadrados com nome iniciando com "ft-").
Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW. *

Sua resposta

08. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 4, como a inclusão de informações ou sua organização?

Sua resposta

Figura 5.62: Questões sobre o recorte 4 da ontologia

A questão 07 descreve o recorte 4 e solicita a opinião sobre sua utilidade para a análise de dados do DW. Por ser a terceira apresentação de uma visualização gráfica, os usuários começaram a se repetir nos comentários ou citar respostas anteriores. Foi citado que a representação é importante para fazer uma conexão correta entre as dimensões e fatos existentes e que permite um entendimento melhor e mais rápido das inter-relações. No entanto, foi citado também que facilitaria sua utilização pelo usuário se oferecesse uma interface mais intuitiva e que a apresentação das setas auxilia o entendimento, mas passa uma sensação de insegurança quanto ao correto sentido das mesmas. É possível que essa insegurança ocorreu pela falta de uma legenda com o significado da relação.

A questão 08 busca obter críticas ou sugestões sobre a representação de conhecimento apresentada no recorte 4. Para este recorte, foi sugerido que fossem informados mais detalhes, como qual dado é hierarquicamente superior, a relação de dependência dos dados, e o domínio.

5.4.4 Considerações sobre os resultados

Consolidando as informações obtidas com a reunião realizada com os usuários, as opiniões sobre uma representação visual para apoiar a tarefa de análise e as respostas sobre os recortes de ontologia apresentados, conclui-se que, na visão dos usuários participantes da pesquisa, uma representação do conhecimento é útil, importante e pode apoiar a análise dos dados do DW.

Foram feitas também algumas sugestões de enriquecimento da representação, como a inclusão de legendas e de mais informações relativas às instâncias. Algumas delas dependem de evoluções no OntoDW, mas outras estão associadas ao Protégé. Apesar de gerar visualizações de boa qualidade gráfica, a ferramenta Protégé se apresenta como de difícil manipulação, até para analistas de TI. Desta forma, ele não está apto para utilização pelo usuário de negócio para consulta aos conceitos devido à dificuldade de uso. Com uma ferramenta de melhor usabilidade, a ontologia gerada poderia permitir a seleção de métricas que se quer analisar e as dimensões que elas teriam em comum para uma representação da análise desejada. Dessa forma, seria fácil a seleção de outras métricas para análise conjunta com o valor do salário.

6 – Trabalhos relacionados

Este capítulo apresenta os trabalhos relacionados com o presente projeto e as principais diferenças existentes.

Foi realizado levantamento bibliográfico à procura por trabalhos relacionados ao problema em questão, e foram encontrados alguns estudos que já exploraram a geração de ontologias a partir de estrutura de dados. Apesar de existirem na literatura diversas abordagens de geração automática de ontologias [Prat, Akoka and Comyn-Wattiau, 2012] [Prat, Megdiche and Akoka, 2012] [Dou, Qin and Lependu, 2010], tais abordagens requerem a existência de outras fontes de dados externas ao sistema que se quer obter um modelo de representação de conhecimento, como modelos de dados e ontologias para alinhamento.

Prat et al. [Prat, Akoka and Comyn-Wattiau, 2012] [Prat, Megdiche and Akoka, 2012] abordam a geração de ontologia OWL-DL a partir de um modelo de dados multi-dimensional. Eles, entretanto, têm como premissa a existência de um modelo de dados conceitual, que representa uma enorme limitação para sua aplicabilidade na prática. Além disso, o conjunto de regras do OntoDW difere das regras definidas por Prat et al. [Prat, Akoka and Comyn-Wattiau, 2012] [Prat, Megdiche and Akoka, 2012] porque elas não utilizam os dados e metadados do DW como elementos de entrada, somente definições de modelos lógicos, e as regras deles não inferem conceitos de sistemas de BI, somente mapeiam conceitos já identificados no modelo lógico para conceitos na ontologia de saída.

Gil et al. [Gil and Martin-Bautista, 2014] [Gil, Martín-Bautista and Contreras, 2010] apresentam uma metodologia para aprendizado de ontologia (SMOL) composta de fases ao longo de um processo estruturado, utilizando fontes de dados heterogêneas (bancos de dados, ontologias e textos). Entretanto, técnicas ou métodos para a geração da ontologia e as etapas do processo não são descritas. Com isso, não é comprovada a capacidade da metodologia em cumprir o que propõe.

[El Idrissi, Baïna and Baïna, 2013] apresentam um levantamento prático de métodos que utilizam estruturas de bancos de dados como entrada para o processo de aprendizado de ontologia. Os autores concluem, a partir deste levantamento, que não existe ferramenta que extraia automaticamente uma ontologia de aplicação a partir de estrutura de banco de dados, o que reforça a motivação para a estratégia proposta no presente trabalho.

[Dou, Qin and Lependu, 2010] propõem um framework para a descoberta automática de mapeamentos entre esquemas de bancos de dados e ontologias, e um algoritmo de tradução de consultas, mas não possibilitam a geração de ontologia com os conceitos da aplicação. Este framework, com a utilização de diferentes ontologias e esquemas e os dados associados a eles, será capaz de extrair um conjunto de regras de mapeamento de primeira ordem que descrevem como as ontologias e esquemas de entrada se relacionam entre si. Portanto, espera-se que exista uma ontologia inicial do sistema para gerar uma ontologia saída.

Moreira et. al [Moreira et. al, 2014] [Moreira et. al, 2015] apresentam uma abordagem ontológica para a derivação de esquemas multidimensionais, usando categorias de uma ontologia de fundamentação (FO) para analisar os domínios de fontes de dados como uma ontologia bem fundamentada. Inicialmente, uma ontologia de domínio é criada e essa ontologia é derivada para um esquema de banco de dados. Esta abordagem tem duas características que não permitem o uso na solução apresentada na Seção 3. A primeira característica é que a abordagem inclui apenas os conceitos de modelagem multidimensional, deixando de fora os conceitos de aplicações OLAP. A segunda característica é que a geração das tabelas multidimensionais no esquema do banco de dados é sempre realizada utilizando as mesmas técnicas. Para utilizar essa abordagem para o processo reverso de geração da ontologia a partir do esquema de banco de dados (objetivo deste trabalho), é necessário que a abordagem cubra técnicas presentes nos modelos do tipo estrela e do tipo floco de neve.

[Boumlik and Bahaj, 2016] apresenta uma abordagem automática para geração de ontologia a partir de um banco de dados relacional, baseada em um conjunto de regras que extraem semântica deste banco de dados e a transforma em um arquivo OWL. Segundo os autores, todas as abordagens existentes para mapeamento de ontologias a partir de bancos de dados relacionais utilizam mapeamento de esquema para transfor-

mar os componentes do modelo de dados conceitual ou modelo físico em conceitos e relações de ontologia. Entretanto, sua abordagem também utiliza técnicas de análise dos dados para detectar os relacionamentos existentes no banco de dados e realiza o mapeamento dos dados para a ontologia. Por outro lado, esta abordagem não extrai semântica de negócio ou de alguma tarefa do banco de dados, realizando apenas uma forma de conversão do esquema de dados para uma ontologia OWL.

Tabela 6.1: Trabalhos relacionados

| Autores | Conceitos | Observações |
|--|--|---|
| [Prat, Akoka and Comyn-Wattiau, 2012] [Prat, Megdiche and Akoka, 2012] | <ul style="list-style-type: none"> • Conjunto de regras de mapeamento • Aborda modelagem multidimensional • Abordagem automática | <ul style="list-style-type: none"> • Não utiliza o DW como fonte de dados • Obrigatória a existência de um modelo conceitual de dados • Não infere conceitos de BI |
| [Gil and Martin-Bautista, 2014] [Gil, Martín-Bautista and Contreras, 2014] | <ul style="list-style-type: none"> • Metodologia para aprendizado de ontologia • Fontes de dados heterogêneas | <ul style="list-style-type: none"> • Não são descritas as técnicas ou métodos utilizados • Não infere conceitos de BI |
| [El Idrissi, Baïna and Baïna, 2013] | <ul style="list-style-type: none"> • Levantamento de métodos para aprendizado de ontologias a partir de bancos de dados | <ul style="list-style-type: none"> • Conclusão de que não existem ferramentas que extraem automaticamente ontologia de aplicação |
| [Dou, Qin and Lependu, 2010] | <ul style="list-style-type: none"> • Framework para descoberta de mapeamentos entre esquemas de banco de dados e ontologias • Abordagem automática | <ul style="list-style-type: none"> • Espera que exista uma ontologia inicial para realizar os mapeamentos • Não infere conceitos de BI |
| [Moreira et. al, 2014] [Moreira et. al, 2015] | <ul style="list-style-type: none"> • Abordagem ontológica para derivação de esquemas multidimensionais | <ul style="list-style-type: none"> • Não aborda conceitos de operações OLAP • Não aborda diferentes técnicas de implementação dos esquemas |
| [Boumlik and Bahaj, 2016] | <ul style="list-style-type: none"> • Conjunto de regras de mapeamento • Utiliza dados e metadados • Abordagem automática | <ul style="list-style-type: none"> • Não infere conceitos do banco de dados • Não aborda conceitos de BI |
| [Vieira, Tanaka and Moura, 2003] | <ul style="list-style-type: none"> • Metodologia e heurística para conversão de esquema de dados • Utiliza dados e metadados • Abordagem automática | <ul style="list-style-type: none"> • Não infere conceitos do banco de dados • Não aborda conceitos de BI |

[Vieira, Tanaka and Moura, 2003] apresenta uma metodologia para engenharia reversa e uma heurística para extração de um modelo conceitual a partir de um esquema de dados relacional. Foi também implementada uma ferramenta com a finalidade de apoiar os usuários no processo de construção de ontologias, geradas de forma automática nas linguagens DAML+OIL ou XTM. O objetivo da proposta é realizar a extração de esquemas de bancos de dados para facilitar sua geração, disponibilização e publicação na Web. Desta forma, esta abordagem não extrai semântica de negócio ou de alguma

tarefa do banco de dados, realizando apenas uma forma de conversão do esquema de dados para uma ontologia.

A análise dos trabalhos consolidados na Tabela 6.1 mostrou a ausência de soluções para a geração de ontologias de aplicação a partir de DWs de forma automática. Em particular, a utilização de uma outra fonte de dados que não uma estrutura de dados multidimensional para a geração de uma ontologia demandaria a existência de documentação atualizada em sincronia com os conceitos implementados ao longo do ciclo de vida do sistema, que não é realístico na prática. É muito difícil manter uma outra fonte de informação disponível para utilização na geração de uma ontologia que esteja sempre atualizada. Por outro lado, em um sistema de BI baseado num DW, a estrutura de dados multidimensional é parte do sistema implementado.

7 – Conclusão

Este capítulo apresenta os principais resultados deste trabalho e aponta suas contribuições e possibilidades de trabalhos futuros para a continuidade da pesquisa.

Este trabalho apresentou um conjunto de regras de mapeamento para gerar automaticamente uma ontologia para sistemas de BI a partir de *Data Warehouses*, contribuindo para resolver a falta de uma representação do conhecimento formal que explicitamente e semanticamente descreva os dados e metadados de sistemas de BI armazenados no DW.

Como vantagem, o uso de elementos DW para gerar uma ontologia fornece uma fonte de informação compartilhada com o sistema de BI, garantindo o alinhamento entre os conceitos implementados no sistema e os conceitos estruturais que a ontologia extraída se propõe a representar. Além disso, esta fonte de informação permite a inferência de conceitos do domínio de BI, tais como agregabilidades, tarefa mais difícil de executar usando um banco de dados operacional. As características dos dados armazenados, tais como volume e esparsidade, também podem ser utilizadas para inferir os elementos da ontologia. Por exemplo, uma tabela agregada de empregados por faixa etária tende a ser menos volumosa que uma tabela de fato no nível de empregado ou idade.

A geração de ontologias a partir de DWs apresenta como desafios algumas questões que são características inerentes comumente encontradas em sistemas de BI, como o grande volume de dados armazenados em estruturas de dados, que torna difícil a manipulação dos dados armazenados no repositório e a estrutura que os contém, e a desnormalização de modelos de dados que torna difícil identificar a relação entre as classes e suas propriedades.

Esta geração da ontologia deve ser automática devido a problemas relacionados com a sua construção manual. Isto facilita manter a consistência da ontologia com os elementos DW ao longo do ciclo de vida da aplicação. Em casos de alterações devido a manutenções evolutivas do sistema, quando novos fatos e dimensões são frequentemente incluídos, a abordagem proposta pode ser reexecutada, de modo a atualizar a conceitualização existente.

As regras de mapeamento que compõem a abordagem OntoDW proposta estendem o estado da arte na geração de ontologias a partir de ambientes de BI. Estas regras tratam de aspectos mais específicos de modelagem multidimensional e levam em consideração tanto os dados quanto os metadados presentes nas estruturas de dados do DW.

As contribuições desta proposta são a criação e melhoria das regras de mapeamento entre elementos de *Data Warehouses* e conceitos de ontologias de aplicação, levando em consideração aspectos específicos de modelagem multidimensional e aplicações OLAP para utilizar os dados e metadados no DW, e implementação de uma ferramenta para a geração automática de ontologias, utilizando regras de mapeamento, as informações de domínio do sistema e um metamodelo de tarefa OLAP, além do *Data Warehouse*.

Entretanto, é importante esclarecer que a ontologia gerada não é um resultado pronto para utilização pelos usuários de negócio. Adicionalmente, poderia ser utilizada por analistas de TI para auxílio na identificação de construtos implementados fora da padronização definida ou divergente do que foi especificado pelas áreas de negócio.

Uma das limitações deste trabalho foi o público-alvo de respondentes das pesquisas. Para o público especialista da área de *Business Intelligence*, as dificuldades estão associadas à baixa quantidade de profissionais do mercado brasileiro, em relação a outras sub-áreas de tecnologia da informação, e à complexidade do assunto e tempo necessário envolvidos nas respostas da pesquisa. Por estas questões, o número de respondentes foi considerado suficiente. Para o público de negócio, o baixo número de respondentes se deve ao foco nos usuários do DW objeto do estudo, para dar maior assertividade às respostas. Apesar do baixo número de pessoas, o envolvimento obtido resultou em um resultado de boa qualidade.

Outra limitação do trabalho foi em relação ao acesso à organização para a execução do estudo de caso. Por este motivo, as tecnologias envolvidas na construção da ferramenta desenvolvida foram escolhidas não somente pelo conhecimento do autor, mas também para tornar o código mais portátil, de mais fácil aplicação no ambiente do estudo de caso e de execução.

Como trabalhos futuros que possibilitem dar continuidade a esta pesquisa, são sugeridos:

- Enriquecimento do conjunto de regras de mapeamento para geração de axiomais na ontologia OWL, com o intuito de identificar mais relações entre os conceitos identificados;
- Melhoria das regras definidas, buscando abranger outras possibilidades de implementação de esquemas de dados multidimensionais que podem gerar resultados diferentes se tiverem regras mais específicas. Como exemplos, podem ser citadas a ocorrência de múltiplas relações entre 2 dimensões e dimensões implementadas somente em colunas de tabela de fato, sem a existência de uma tabela de dimensão.
- Utilização de outras técnicas para definição de nomes dos conceitos identificados, como processamento de linguagem natural em documentos da organização, por exemplo;
- Evolução do metamodelo de tarefa OLAP para atender a mais necessidades do usuário. Como exemplos, podemos citar a criação de propriedades para explicitar a relação de superioridade entre os *roll ups* em uma hierarquia e para a qualificação de dimensões por tipo;
- Criação de ferramenta que utilize a ontologia para prover ao usuário uma análise guiada dos dados, juntamente com o histórico de utilização de análises e preferências. Essa ferramenta também poderia permitir a realização de ajustes na ontologia para correção de eventuais imprecisões e também prover o cruzamento com dados de outras fontes de dados, como planilhas e outros esquemas de dados relacionais. Desta forma, a abordagem poderia realizar a geração automática ou semi-automática da ontologia, conforme for mais útil para o usuário da ferramenta;

- Pesquisas relacionadas a Gestão do Conhecimento, com a integração dos conceitos da ontologia OWL gerada pelo OntoDW com conceitos formalmente definidos, sejam de negócio ou de outros sistemas.

Referências

- Airinei, D., and Homocianu, D. "DSS vs. business intelligence." *Revista Economica*, 2009.
- Andoh-Baidoo, et. al. "Business Intelligence & Analytics Education: An Exploratory Study of Business & Non-Business School IS Program Offerings." *Twentieth Americas Conference on Information Systems*, Savannah, Georgia, 2014.
- Boumlik, A., and Bahaj, M (2016). "Advanced Set of Rules to Generate Ontology from Relational Database." *JSW 11(1)*, pp. 27-43, 2016.
- Breitman, K., Casanova, M. A., and Truszkowski, W. *Semantic web: concepts, technologies and applications*. Springer Science & Business Media, 2007.
- Chaudhuri, S., Dayal, U., and Narasayya, V. "An overview of business intelligence technology." *Communications of the ACM* 54.8, 2011, pp. 88-98.
- Dou, D., Qin, H., and Lependu, P. "OntoGrate: Towards automatic integration for relational databases and the semantic web through an ontology-based framework." *International Journal of Semantic Computing* 4.01, 2010, pp. 123-151.
- El Idrissi, B., Baïna, S., and Baïna, K. "Automatic generation of ontology from data models: a practical evaluation of existing approaches." *IEEE 7th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 2013.
- Euzenat, J., and Shvaiko, P. *Ontology matching*. Heidelberg: Springer, 2013.
- Gil, R., and Martín-Bautista, M. J. "SMOL: a systemic methodology for ontology learning from heterogeneous sources." *Journal of Intelligent Information Systems* 42.3, 2014, pp. 415-455.
- Gil, R., Martín-Bautista, M. J., and Contreras, L. "Applying an ontology learning methodology to a relational database: University case study." *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*. IEEE, 2010.
- Gruber, T. "A translation approach to portable ontology specifications." *Knowledge acquisition* 5.2, 1993, pp. 199-220.
- Guarino, N. "Semantic matching: Formal ontological distinctions for information organization, extraction, and integration." *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*. Springer Berlin Heidelberg, 1997. pp. 139-170.
- Inmon, W. *Building the data warehouse, 3rd ed*. Wiley Computer Publishing, 428p., 2002.
- Kimball, R. "A dimensional modeling manifesto." *DBMS* 10.9, 1997, pp. 58-70.
- Luhn, H. P. "A business intelligence system." *IBM Journal of Research and Development* 2.4, 1958, pp. 314-319.

- Mansingh, G., Osei-Bryson, K. M., and Reichgelt, H. "Using ontologies to facilitate post-processing of association rules by domain experts." *Information Sciences* 181.3 2011, pp. 419-434.
- Melchert, F., Winter, R., and Klesse, M. "Aligning process automation and business intelligence to support corporate performance management." *AMCIS 2004 Proceedings*, 2004, p. 507.
- Moreira, J., et al. "OntoWarehousing—multidimensional design supported by a foundational ontology: a temporal perspective." *International Conference on Data Warehousing and Knowledge Discovery*. Springer International Publishing, 2014.
- Moreira, J., et al. "Hybrid Multidimensional Design for Heterogeneous Data Supported by Ontological Analysis: an Application Case in the Brazilian Electric System Operation." *EDBT/ICDT Workshops*. 2015.
- Negash, S. "Business Intelligence." *Communications of the Association for Information Systems*, 13, 2004.
- Power, D. J. "Decision support systems: a historical overview." *Handbook on Decision Support Systems 1*. Springer Berlin Heidelberg, 2008. pp. 121-140.
- Prat, N., Akoka, J., and Comyn-Wattiau, I. "Transforming multidimensional models into OWL-DL ontologies." *2012 Sixth International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 2012.
- Prat, N., Megdiche, I., and Akoka, J. "Multidimensional models meet the semantic web: defining and reasoning on OWL-DL ontologies for OLAP." *Proceedings of the fifteenth international workshop on Data warehousing and OLAP*. ACM, 2012.
- Ross, M. "Differences of Opinion—the Kimball Bus Architecture and the Corporate Information Factory: What are the Fundamental Differences." 2004.
- Sell, D., et al. "Adding Semantics to Business Intelligence: Towards a Smarter Generation of Analytical Tools." *Business Intelligence—Solution for Business Development*, p. 33, 2011.
- Sidorova, A., and Torres, R. "Business Intelligence and Analytics: A Capabilities Dynamization View." *Twentieth Americas Conference on Information Systems*, Savannah, Georgia, 2014.
- Staab, S., and Studer, R. *Handbook on ontologies*. Springer Science & Business Media, 2013.
- Vieira, A. A., Tanaka, A. K., and Moura, A. M. D. C. "Ferramenta para Extração de Ontologias a Partir de Bancos de Dados Relacionais." 2003.
- W3C. "OWL Web Ontology Language Guide". Disponível em: <https://www.w3.org/TR/2004/REC-owl-guide-20040210/#OwlVarieties>. Acesso em 01 Mai 2016.
- Wuensch, K. L. "What is a Likert Scale? and How Do You Pronounce Likert?." *East Carolina University*, 2005.

ANEXO I – Questionário 1 para profissionais de BI do mercado

Pesquisa - Modelos conceituais de Data Warehouses

Introdução

Esta pesquisa é parte de um projeto de pesquisa de Mestrado elaborado na UNIRIO (PPGI) sobre a geração de modelos conceituais.

O presente estudo é composto por:

- (1) conjunto de perguntas sobre o conhecimento e a experiência do entrevistado com modelagem dimensional e conceitos de aplicações de Business Intelligence (BI);
- (2) um conjunto de fragmentos do modelo físico de um Data Warehouse (DW);
- (3) um conjunto de conceitos de BI identificados no Data Warehouse;

Recomendações para o preenchimento do questionário:

- Preencha o questionário sequencialmente, respondendo as perguntas na ordem em que aparecem.

Informações adicionais:

- Suas respostas serão consideradas na pesquisa e contribuirão para publicações relevantes. Sinta-se à vontade para responder com sinceridade.
- O preenchimento do questionário é anônimo.
- Tempo estimado para preenchimento: 10 (dez) minutos.

Muito Obrigado pela sua participação!

Autor: Tiago Outerelo da Silva

PRÓXIMA

 Página 1 de 8

Análise de Perfil

Algumas informações de perfil são necessárias para a análise e comparação dos resultados. Por favor, responda as questões abaixo de acordo com a escala apresentada:

Qual seu domínio nos conceitos de modelagem multidimensional? *

| | 1 | 2 | 3 | 4 | 5 | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual seu domínio na implementação de modelos multidimensionais? *

| | 1 | 2 | 3 | 4 | 5 | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual sua experiência, em anos, na implementação de modelos multidimensionais? *

Sua resposta

Qual seu domínio nos conceitos de aplicações OLAP (ex.: roll up, hierarquia, atributo, métrica)? *

| | 1 | 2 | 3 | 4 | 5 | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual seu domínio na implementação de aplicações OLAP? *


| | 1 | 2 | 3 | 4 | 5 | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual sua experiência, em anos, na implementação de aplicações OLAP? *

Sua resposta

VOLTAR




PRÓXIMA

 Página 2 de 8

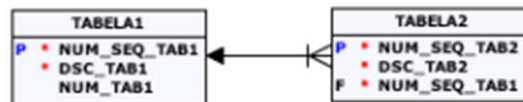
Avaliação dos conceitos representados no DW

A seguir é apresentada a notação utilizada nos fragmentos que serão apresentados do Data Warehouse.

Notação

| Componente | Notação | | |
|----------------------------|--|-------------|-----------------|
| Tabela | <table border="1"><tr><td>NOME_TABELA</td></tr><tr><td>P * NOME_COLUNA</td></tr></table> | NOME_TABELA | P * NOME_COLUNA |
| NOME_TABELA | | | |
| P * NOME_COLUNA | | | |
| Relação 1 para muitos |  | | |
| Relação 0 para muitos |  | | |
| Relação muitos para muitos |  | | |
| Chave primária | P | | |
| Chave estrangeira | F | | |
| Coluna não nula | * | | |

Exemplo de fragmento



VOLTAR

PRÓXIMA

Página 3 de 8

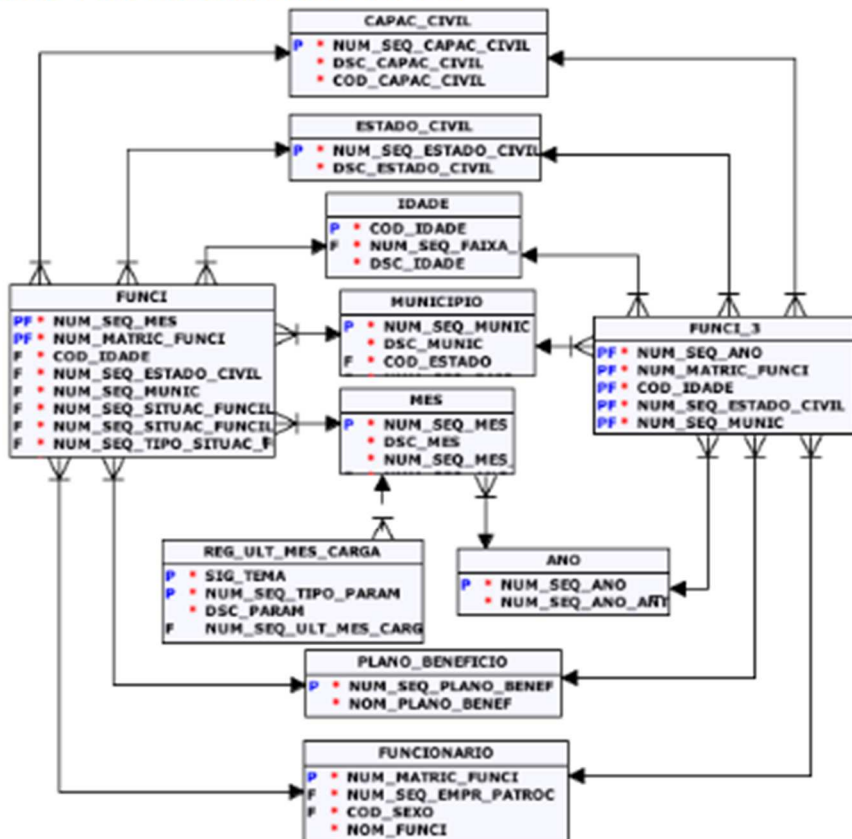
Avaliação dos conceitos representados no DW

No âmbito desta pesquisa, fatos são as tabelas que armazenam os indicadores para análise e dimensões são as tabelas que armazenam os agrupamentos/categorizações desses indicadores.

Ex.: Em um DW de vendas, uma tabela de fato poderia armazenar a quantidade vendida por produto (tabela VENDAS) e existiria uma tabela de dimensão com os produtos vendidos (tabela PRODUTO).

Com base nestas definições, por favor responda as questões apresentadas abaixo sobre o Recorte 1:

Recorte 1 de modelo físico de DW




01. Informe qual conceito de modelagem dimensional (dimensão ou fato/agregação) corresponde a cada tabela do Recorte 1, se for o caso: *

| | DIMENSÃO | FATO/AGREGAÇÃO | NENHUM DOS DOIS |
|-------------------|-----------------------|-----------------------|-----------------------|
| ANO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| CAPAC_CIVIL | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| ESTADO_CIVIL | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCI | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCIONARIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCI_3 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| IDADE | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| MES | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| MUNICIPIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| PLANO_BENEFICIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| REG_ULT_MES_CARGA | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

VOLTAR

PRÓXIMA

 Página 4 de 8

Avaliação dos conceitos representados no DW

Em esquemas de dados multidimensionais, é comum a implementação de mais de um agrupamento/categorização em uma mesma tabela (dimensão) para prover melhor desempenho.

No âmbito desta pesquisa, chamaremos esse agrupamento/categorização de nível de dimensão. Uma dimensão contém 1 ou mais níveis de dimensão. Cada coluna de uma tabela de dimensão pertence a 1 nível de dimensão (apenas a 1). Cada coluna de um nível de dimensão será chamada de atributo.

Ex.: Em um DW de vendas, uma possível dimensão "Produto" poderia armazenar um nível de dimensão do produto (com as colunas "codigo_produto" e "nome_produto", sendo o menor grão da dimensão), um nível com a cor do produto (com a coluna "nome_cor_produto") e outro nível com a categoria do produto (com a coluna "nome_categoria_produto").

Importante observar que toda dimensão tem pelo menos 1 nível, que é o mesmo da PK da tabela, representando o menor "grão" da tabela. No exemplo acima, o nível que identifique cada produto terá pelo menos as colunas que compõem a PK, podendo incluir outras colunas que qualifiquem o produto, como o nome do produto.

Com base nestas definições, por favor responda as questões apresentadas abaixo sobre o Recorte 2:

Recorte 2 de modelo físico de DW

| FUNCIONARIO ▲ | |
|---------------|--------------------|
| P | NUM_MATRIC_FUNC1 |
| F | COD_SEXO |
| | NOM_FUNC1 |
| | DAT_NASC_FUNC1 |
| | NUM_CPF_FUNC1 |
| F | NUM_SEQ_DIA_APOSE |
| F | NUM_SEQ_DIA_EXONDO |
| F | NUM_SEQ_DIA_FALECI |

02. Considerando que a tabela FUNCIONARIO apresentada no Recorte 2 é uma dimensão, informe o número de níveis de dimensão que ela contém: *

Sua resposta

03. Baseado nos níveis de dimensão encontrados na tabela do Recorte 2, informe o número de operações de Roll Up (mudança de grão para um nível hierarquicamente superior) que podem ser realizadas: *

Sua resposta

04. Caso tenha encontrado Roll Up's na tabela do Recorte 2, eles poderiam ser agrupados em uma hierarquia?: *

Sim

Não


05. Cada coluna da matriz abaixo representa um possível nível de dimensão encontrado. Agrupe atributos (colunas) que sejam de um mesmo nível de dimensão a uma mesma coluna. Ex.: Caso tenha encontrado apenas 1 nível de dimensão, marque a 1ª coluna em todas as linhas: *

Atributos (Colunas da tabela) x Níveis de dimensão

| | Nível 1 | Nível 2 | Nível 3 | Nível 4 | Nível 5 | Nível 6 |
|--------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| NUM_MATRIC_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| COD_SEXO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NOM_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| DAT_NASC_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_CPF_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_APOSE | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_EXONDO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_FALECI | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

VOLTAR

PRÓXIMA

 Página 5 de 8

Avaliação dos conceitos representados no DW

Chamamos de Roll Up a operação de mudança de um determinado nível de dimensão numa análise para outro nível de dimensão menos detalhado. Esses níveis de dimensão podem estar numa mesma tabela (dimensão) ou em tabelas (dimensões) diferentes.

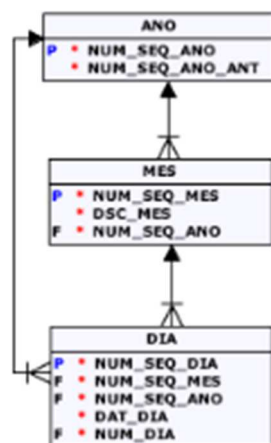
Ex.: Em um DW de vendas, ao realizar a análise das vendas no nível de produto (origem), é possível realizar um Roll Up e consolidar as vendas para o nível de categoria (destino), que estão na mesma dimensão "Produto". Caso a dimensão "Produto" esteja relacionada a uma outra dimensão chamada "Tipo_Produto", um Roll Out do nível de produto (origem) para o nível de tipo de produto (destino) também seria possível.

Uma hierarquia é conjunto de Roll Up's relacionados, onde o destino de um Roll Up é a origem de outro Roll Up.

Ex.: Para o DW de vendas, teríamos uma hierarquia com os Roll Up's da dimensão "Produto" e outra hierarquia com o Roll Up entre as dimensões "Produto" e "Tipo_Produto".

Com base nestas definições, por favor responda as questões apresentadas abaixo sobre o Recorte 3:

Recorte 3 de modelo físico de DW



06. Considerando que as tabelas apresentadas no Recorte 3 são dimensões, informe o número de operações possíveis de Roll Up que podem ser realizadas entre elas: *

Sua resposta

07. Caso tenha encontrado Roll Up's na tabela do Recorte 3, eles poderiam ser agrupados em uma hierarquia?: *

- Sim
- Não

VOLTAR

PRÓXIMA

Página 6 de 8

Avaliação dos conceitos representados no DW

No âmbito deste trabalho, chamaremos de medidas os indicadores que se quer analisar sobre determinado assunto. As medidas são representadas num Data Warehouse como colunas nas tabelas de fato.

Ex.: Em um DW de vendas, o fato "Vendas" poderia conter as medidas de quantidade vendida e de valor da venda.

Com base nestas definições, por favor responda as questões abaixo sobre o Recorte 4.

Considere que suas 3 tabelas apresentadas são fatos:

Recorte 4 de modelo físico de DW

| FUNC1_3 | FUNC1 | REG_ULT_MES_CARGA |
|------------------------------|------------------------------|---------------------------|
| PF * NUM_SEQ_ANO | PF * NUM_SEQ_MES | P * SIG_TEMA |
| PF * NUM_MATRIC_FUNC1 | PF * NUM_MATRIC_FUNC1 | P * NUM_SEQ_TIPO_PARAM |
| PF * COD_IDADE | P * COD_IDADE | P * DSC_PARAM |
| PF * NUM_SEQ_ESTADO_CIVIL | P * NUM_SEQ_ESTADO_CIVIL | F * NUM_SEQ_ULT_MES_CARGA |
| PF * NUM_SEQ_MUNIC | P * NUM_SEQ_MUNIC | |
| PF * NUM_SEQ_SITUAC_FUNC1 | P * NUM_SEQ_PLANO_BENEF | |
| PF * NUM_SEQ_PLANO_BENEF | P * NUM_SEQ_PERIODO_NORMAT | |
| PF * NUM_SEQ_PERIODO_NORMAT | P * NUM_SEQ_CAPAC_CIVIL | |
| PF * NUM_SEQ_CAPAC_CIVIL | * QTD_FUNC1 | |
| P * IND_RECSTO_APOSE | * QTD_DEP | |
| * QTD_FUNC1 | * QTD_PENSTA | |
| * QTD_BENEF | * QTD_DESPIL | |
| * QTD_PENSTA | * VAL_SALAR_PARTIC | |
| * VAL_SALAR_PARTIC | * NUM_DIA_CONTRI_INSS_FORA | |
| * NUM_DIA_CONTRI_INSS_PATROC | * NUM_DIA_CONTRI_INSS_PATROC | |
| | * IND_RECSTO_APOSE | |

08. Informe a quantidade de medidas presentes na tabela FUNC1_3: *

Sua resposta

09. Informe a quantidade de medidas presentes na tabela FUNC1: *

Sua resposta

10. Informe a quantidade de medidas presentes na tabela REG_ULT_MES_CARGA: *

Sua resposta

11. Caso tenha encontrado medidas nas tabelas do Recorte 4, informe quantas dessas medidas estão presentes em mais de uma tabela: *

Sua resposta

VOLTAR

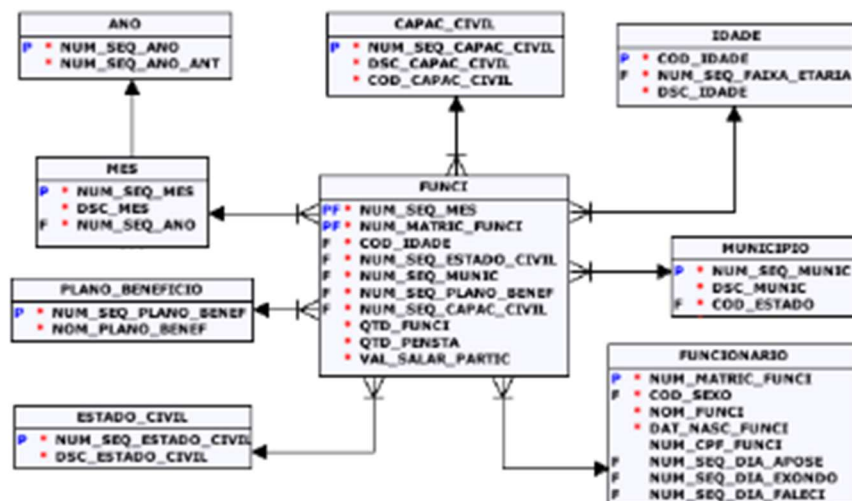
PRÓXIMA

Página 7 de 8

Avaliação dos conceitos representados no DW

Por favor, responda as questões apresentadas abaixo sobre o Recorte 5:

Recorte 5 de modelo físico de DW



12. Caso encontre medidas nas tabelas do Recorte 5, informe a quantidade de dimensões distintas que podem ser usadas para analisar essas medidas: *

Sua resposta

13. Caso encontre medidas nas tabelas do Recorte 5, informe a quantidade de hierarquias distintas que podem ser usadas para analisar essas medidas: *

Sua resposta

VOLTAR

ENVIAR

Página 8 de 8

ANEXO II – Questionário 2 para profissionais de BI da organização do estudo de caso

Pesquisa - Modelos conceituais de Data Warehouses

Introdução

Esta pesquisa é parte de um projeto de pesquisa de Mestrado elaborado na UNIRIO (PPGI) sobre a geração de modelos conceituais.

O presente estudo é composto por:

- (1) conjunto de perguntas sobre o conhecimento e a experiência do entrevistado com modelagem dimensional e conceitos de aplicações de Business Intelligence (BI);
- (2) um conjunto de fragmentos do modelo físico de um Data Warehouse (DW);
- (3) um conjunto de conceitos de BI identificados no Data Warehouse;

Recomendações para o preenchimento do questionário:

- Preencha o questionário sequencialmente, respondendo as perguntas na ordem em que aparecem.

Informações adicionais:

- Suas respostas serão consideradas na pesquisa e contribuirão para publicações relevantes. Sinta-se à vontade para responder com sinceridade.
- O preenchimento do questionário é anônimo.
- Tempo estimado para preenchimento: 10 (dez) minutos.

Muito Obrigado pela sua participação!

Autor: Tiago Outerelo da Silva

PRÓXIMA

Página 1 de 8

Análise de Perfil

Algumas informações de perfil são necessárias para a análise e comparação dos resultados. Por favor, responda as questões abaixo de acordo com a escala apresentada:

Qual seu domínio nos conceitos de modelagem multidimensional? *

| | 1 | 2 | 3 | 4 | 5 | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual seu domínio na implementação de modelos multidimensionais? *

| | 1 | 2 | 3 | 4 | 5 | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual sua experiência, em anos, na implementação de modelos multidimensionais? *

Sua resposta

Qual seu domínio nos conceitos de aplicações OLAP (ex.: roll up, hierarquia, atributo, métrica)? *

| | 1 | 2 | 3 | 4 | 5 | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual seu domínio na implementação de aplicações OLAP? *

| | 1 | 2 | 3 | 4 | 5 | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual sua experiência, em anos, na implementação de aplicações OLAP? *

Sua resposta

Há quanto tempo trabalha na empresa atual? *

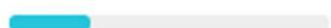
Sua resposta

Qual sua função atualmente? *

Sua resposta

VOLTAR

PRÓXIMA

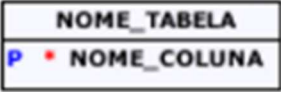





Página 2 de 8

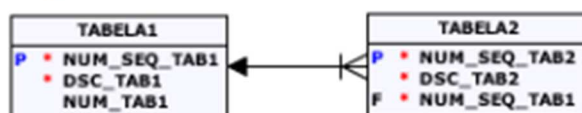
Avaliação dos conceitos representados no DW

A seguir é apresentada a notação utilizada nos fragmentos que serão apresentados do Data Warehouse.

Notação

| Componente | Notação |
|----------------------------|--|
| Tabela |  |
| Relação 1 para muitos |  |
| Relação 0 para muitos |  |
| Relação muitos para muitos |  |
| Chave primária | P |
| Chave estrangeira | F |
| Coluna não nula | * |

Exemplo de fragmento



VOLTAR

PRÓXIMA

Página 3 de 8

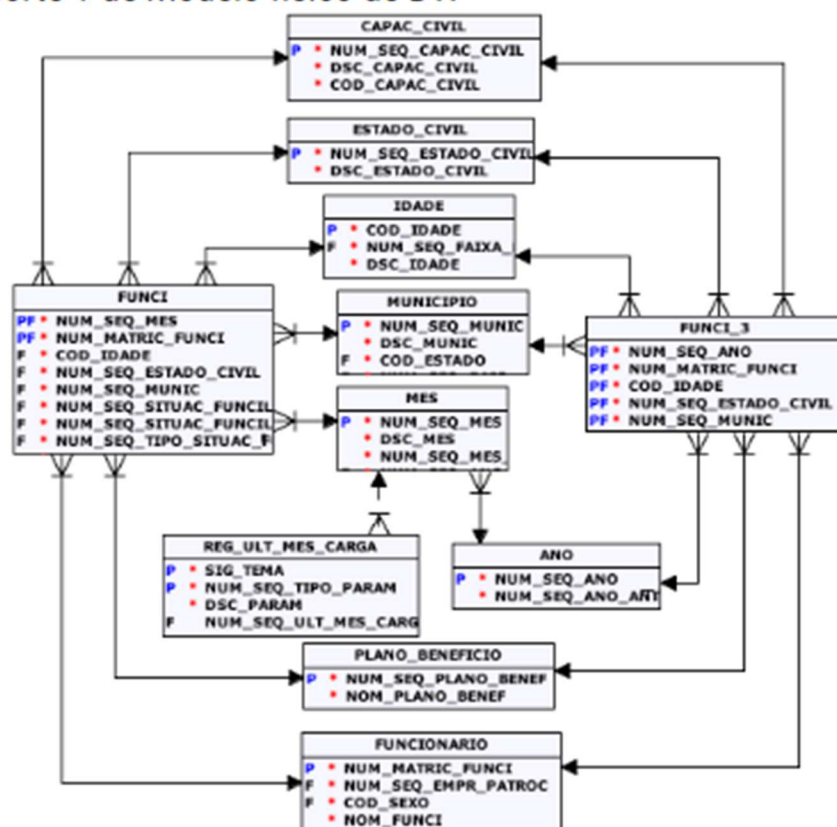
Avaliação dos conceitos representados no DW

No âmbito desta pesquisa, fatos são as tabelas que armazenam os indicadores para análise e dimensões são as tabelas que armazenam os agrupamentos/categorizações desses indicadores.

Ex.: Em um DW de vendas, uma tabela de fato poderia armazenar a quantidade vendida por produto (tabela VENDAS) e existiria uma tabela de dimensão com os produtos vendidos (tabela PRODUTO).

Com base nestas definições, por favor responda as questões apresentadas abaixo sobre o Recorte 1:

Recorte 1 de modelo físico de DW



01. Informe qual conceito de modelagem dimensional (dimensão ou fato/agregação) corresponde a cada tabela do Recorte 1, se for o caso: *

| | DIMENSÃO | FATO/AGREGAÇÃO | NENHUM DOS DOIS |
|-------------------|-----------------------|-----------------------|-----------------------|
| ANO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| CAPAC_CIVIL | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| ESTADO_CIVIL | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCI | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCIONARIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| FUNCI_3 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| IDADE | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| MES | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| MUNICIPIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| PLANO_BENEFICIO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| REG_ULT_MES_CARGA | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

As informações presentes no Recorte 1 foram suficientes para responder a questão acima? De qual informação sentiu falta?


Sua resposta

Comente sobre a dificuldade de responder à questão acima, a estratégia para chegar à resposta ou qualquer outro assunto que achar pertinente.

Sua resposta

VOLTAR

PRÓXIMA

 Página 4 de 8

Avaliação dos conceitos representados no DW

Em esquemas de dados multidimensionais, é comum a implementação de mais de um agrupamento/categorização em uma mesma tabela (dimensão) para prover melhor desempenho.

No âmbito desta pesquisa, chamaremos esse agrupamento/categorização de nível de dimensão. Uma dimensão contém 1 ou mais níveis de dimensão. Cada coluna de uma tabela de dimensão pertence a 1 nível de dimensão (apenas a 1). Cada coluna de um nível de dimensão será chamada de atributo.

Ex.: Em um DW de vendas, uma possível dimensão "Produto" poderia armazenar um nível de dimensão do produto (com as colunas "codigo_produto" e "nome_produto", sendo o menor grão da dimensão), um nível com a cor do produto (com a coluna "nome_cor_produto") e outro nível com a categoria do produto (com a coluna "nome_categoria_produto").

Importante observar que toda dimensão tem pelo menos 1 nível, que é o mesmo da PK da tabela, representando o menor "grão" da tabela. No exemplo acima, o nível que identifique cada produto terá pelo menos as colunas que compõem a PK, podendo incluir outras colunas que qualifiquem o produto, como o nome do produto.

Com base nestas definições, por favor responda as questões apresentadas abaixo sobre o Recorte 2:

Recorte 2 de modelo físico de DW

| FUNCIONARIO ▲ | |
|---------------|--------------------|
| P | NUM_MATRIC_FUNC1 |
| F | COD_SEXO |
| | NOM_FUNC1 |
| | DAT_NASC_FUNC1 |
| | NUM_CPF_FUNC1 |
| F | NUM_SEQ_DIA_APOSE |
| F | NUM_SEQ_DIA_EXONDO |
| F | NUM_SEQ_DIA_FALECI |

02. Considerando que a tabela FUNCIONARIO apresentada no Recorte 2 é uma dimensão, informe o número de níveis de dimensão que ela contém: *

Sua resposta

03. Baseado nos níveis de dimensão encontrados na tabela do Recorte 2, informe o número de operações de Roll Up (mudança de grão para um nível hierarquicamente superior) que podem ser realizadas: *

Sua resposta

04. Caso tenha encontrado Roll Up's na tabela do Recorte 2, eles poderiam ser agrupados em uma hierarquia?: *

Sim

Não

05. Cada coluna da matriz abaixo representa um possível nível de dimensão encontrado. Agrupe atributos (colunas) que sejam de um mesmo nível de dimensão a uma mesma coluna. Ex.: Caso tenha encontrado apenas 1 nível de dimensão, marque a 1ª coluna em todas as linhas: *

Atributos (Colunas da tabela) x Níveis de dimensão

| | Nível 1 | Nível 2 | Nível 3 | Nível 4 | Nível 5 | Nível 6 |
|--------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| NUM_MATRIC_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| COD_SEXO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NOM_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| DAT_NASC_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_CPF_FUNC1 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_APOSE | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_EXONDO | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| NUM_SEQ_DIA_FALECI | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

As informações presentes no Recorte 2 foram suficientes para responder às questões acima? De qual informação sentiu falta?


Sua resposta

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

Sua resposta

VOLTAR

PRÓXIMA

 Página 5 de 8

Avaliação dos conceitos representados no DW

Chamamos de Roll Up a operação de mudança de um determinado nível de dimensão numa análise para outro nível de dimensão menos detalhado. Esses níveis de dimensão podem estar numa mesma tabela (dimensão) ou em tabelas (dimensões) diferentes.

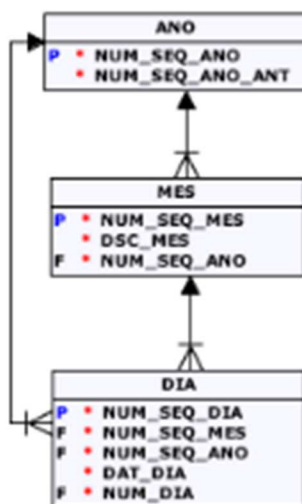
Ex.: Em um DW de vendas, ao realizar a análise das vendas no nível de produto (origem), é possível realizar um Roll Up e consolidar as vendas para o nível de categoria (destino), que estão na mesma dimensão "Produto". Caso a dimensão "Produto" esteja relacionada a uma outra dimensão chamada "Tipo_Produto", um Roll Out do nível de produto (origem) para o nível de tipo de produto (destino) também seria possível.

Uma hierarquia é conjunto de Roll Up's relacionados, onde o destino de um Roll Up é a origem de outro Roll Up.

Ex.: Para o DW de vendas, teríamos uma hierarquia com os Roll Up's da dimensão "Produto" e outra hierarquia com o Roll Up entre as dimensões "Produto" e "Tipo_Produto".

Com base nestas definições, por favor responda as questões apresentadas abaixo sobre o Recorte 3:

Recorte 3 de modelo físico de DW



06. Considerando que as tabelas apresentadas no Recorte 3 são dimensões, informe o número de operações possíveis de Roll Up que podem ser realizadas entre elas: *

Sua resposta

07. Caso tenha encontrado Roll Up's na tabela do Recorte 3, eles poderiam ser agrupados em uma hierarquia?: *

Sim

Não

As informações presentes no Recorte 3 foram suficientes para responder às questões acima? De qual informação sentiu falta?

Sua resposta

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

Sua resposta

VOLTAR

PRÓXIMA



Página 6 de 8

Avaliação dos conceitos representados no DW

No âmbito deste trabalho, chamaremos de medidas os indicadores que se quer analisar sobre determinado assunto. As medidas são representadas num Data Warehouse como colunas nas tabelas de fato.

Ex.: Em um DW de vendas, o fato "Vendas" poderia conter as medidas de quantidade vendida e de valor da venda.

Com base nestas definições, por favor responda as questões abaixo sobre o Recorte 4.

Considere que suas 3 tabelas apresentadas são fatos:

Recorte 4 de modelo físico de DW

| FUNCI_3 | FUNCI | REG_ULT_MES_CARGA |
|------------------------------|------------------------------|---------------------------|
| PF * NUM_SEQ_ANO | PF * NUM_SEQ_MES | P * SIG_TEMA |
| PF * NUM_MATRIC_FUNC1 | PF * NUM_MATRIC_FUNC1 | P * NUM_SEQ_TIPO_PARAM |
| PF * COD_IDADE | F * COD_IDADE | * DSC_PARAM |
| PF * NUM_SEQ_ESTADO_CIVIL | F * NUM_SEQ_ESTADO_CIVIL | F * NUM_SEQ_ULT_MES_CARGA |
| PF * NUM_SEQ_MUNIC | F * NUM_SEQ_MUNIC | |
| PF * NUM_SEQ_SITUAC_FUNC1 | F * NUM_SEQ_PLANO_BENEF | |
| PF * NUM_SEQ_PLANO_BENEF | F * NUM_SEQ_PERIODO_NORMAT | |
| PF * NUM_SEQ_PERIODO_NORMAT | F * NUM_SEQ_CAPAC_CIVIL | |
| PF * NUM_SEQ_CAPAC_CIVIL | * QTD_FUNC1 | |
| P * IND_RECSTO_APOSE | * QTD_DIP | |
| * QTD_FUNC1 | * QTD_PENSA | |
| * QTD_BENEFIC | * QTD_DESPIL | |
| * QTD_PENSA | * VAL_SALAR_PARTIC | |
| * VAL_SALAR_PARTIC | * NUM_DIA_CONTRI_INSS_FORA | |
| * NUM_DIA_CONTRI_INSS_PATROC | * NUM_DIA_CONTRI_INSS_PATROC | |
| | * IND_RECSTO_APOSE | |

08. Informe a quantidade de medidas presentes na tabela FUNCI_3: *

Sua resposta

09. Informe a quantidade de medidas presentes na tabela FUNCI: *

Sua resposta

10. Informe a quantidade de medidas presentes na tabela REG_ULT_MES_CARGA: *

Sua resposta

11. Caso tenha encontrado medidas nas tabelas do Recorte 4, informe quantas dessas medidas estão presentes em mais de uma tabela: *

Sua resposta

As informações presentes no Recorte 4 foram suficientes para responder às questões acima? De qual informação sentiu falta?

Sua resposta

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

Sua resposta

VOLTAR

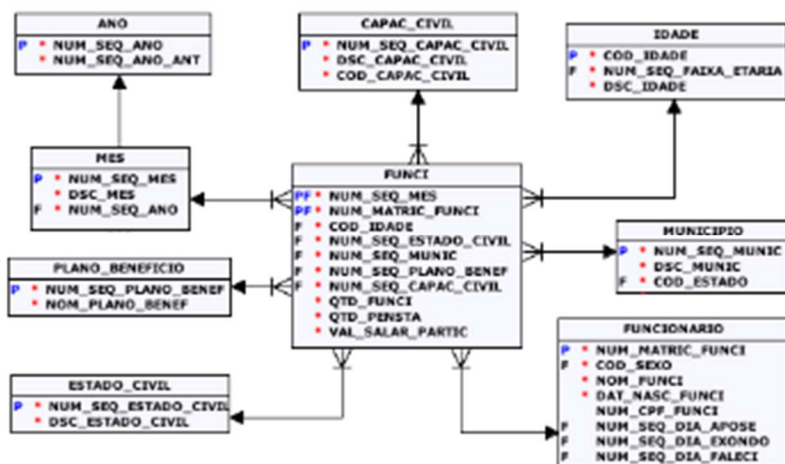
PRÓXIMA

 Página 7 de 8

Avaliação dos conceitos representados no DW

Por favor, responda as questões apresentadas abaixo sobre o Recorte 5:

Recorte 5 de modelo físico de DW



12. Caso encontre medidas nas tabelas do Recorte 5, informe a quantidade de dimensões distintas que podem ser usadas para analisar essas medidas: *

Sua resposta

13. Caso encontre medidas nas tabelas do Recorte 5, informe a quantidade de hierarquias distintas que podem ser usadas para analisar essas medidas: *

Sua resposta

As informações presentes no Recorte 5 foram suficientes para responder às questões acima? De qual informação sentiu falta?

Sua resposta

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

Sua resposta

VOLTAR

ENVIAR

Página 8 de 8

ANEXO III – Questionário 3 para usuários do sistema de BI

Pesquisa - Modelos conceituais de Data Warehouses

Introdução

Esta pesquisa é parte de um projeto de pesquisa de Mestrado elaborado na UNIRIO (PPGI) sobre a geração de modelos conceituais a partir de Data Warehouses.

O presente estudo é composto por:

- (1) conjunto de perguntas sobre a experiência do entrevistado com aplicações de Business Intelligence (BI);
- (2) um conjunto de fragmentos de modelos com conceitos de BI identificados no Data Warehouse (DW);

Recomendações para o preenchimento do questionário:

- Preencha o questionário sequencialmente, respondendo as perguntas na ordem em que aparecem.

Informações adicionais:

- Suas respostas serão consideradas na pesquisa e contribuirão para publicações relevantes. Sinta-se à vontade para responder com sinceridade.
- O preenchimento do questionário é anônimo.
- Tempo estimado para preenchimento: 10 (dez) minutos.

Muito Obrigado pela sua participação!

Autor: Tiago Outerelo da Silva

PRÓXIMA



Página 1 de 7

Análise de Perfil

Algumas informações de perfil são necessárias para a análise e comparação dos resultados.

Lembrando:

- Aplicações de Business Intelligence (BI) oferecem meios para fornecer informações e derivar conhecimento através de ferramentas de análise para tomada de decisão e auxiliam na análise de grandes volumes de dados.

- Data Warehouse é o banco de dados que armazena as informações das aplicações de BI.

Por favor, responda as questões abaixo de acordo com a escala apresentada:

Qual seu domínio nos conceitos de aplicações de BI (ex.: hierarquia, atributo, métrica)? *

| | | | | | | |
|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| | 1 | 2 | 3 | 4 | 5 | |
| Nenhum | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Especialista |

Qual sua experiência, em anos, na utilização de aplicações de BI? *

Sua resposta

Há quanto tempo trabalha na empresa atual? *


Sua resposta

Qual sua função atualmente? *

Sua resposta

VOLTAR

PRÓXIMA

 Página 2 de 7

Avaliação dos conceitos representados no DW

Responda as questões abaixo com base na experiência de utilização de aplicações de BI e nas necessidades e possibilidades de melhoria que identifica nas atividades de trabalho que envolvam análise e geração de informações.

Qual sua opinião sobre a utilidade de uma representação visual que descrevesse as informações implementadas no DW alinhadas com conceitos de BI? Ex.: Demonstrasse as dimensões e as medidas (métricas) disponíveis, as análises possíveis de se realizar. *

Sua resposta

Existiria vantagem entre uma representação gráfica e uma textual? *

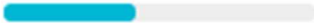
Sua resposta

Se essas informações fossem apresentadas com o uso de termos próprios do negócio ao invés dos nomes no DW, seriam mais úteis? *

Sua resposta

VOLTAR

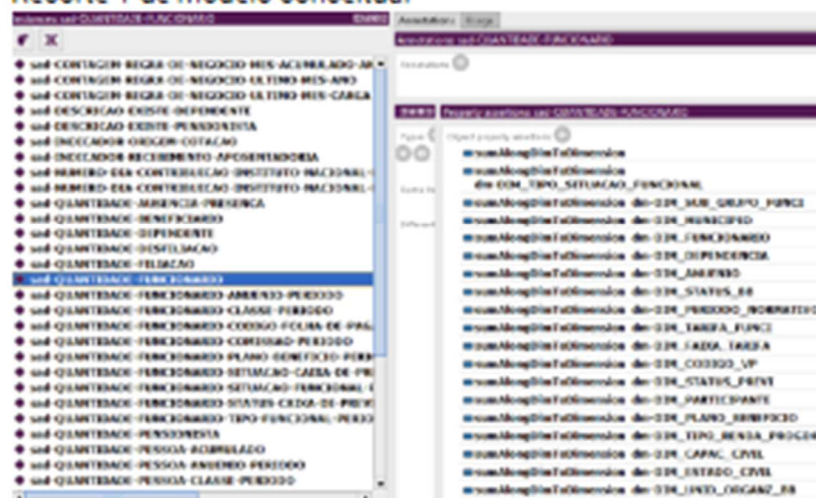
PRÓXIMA

 Página 3 de 7

Avaliação dos conceitos representados no DW

Medidas (ou métricas) são indicadores para análise sobre determinado assunto e dimensões são as tabelas que armazenam os agrupamentos/categorizações desses indicadores. Com base nestas definições, por favor responda as questões apresentadas abaixo sobre o Recorte 1:

Recorte 1 de modelo conceitual



01. O Recorte 1 mostra as dimensões disponíveis para se analisar (lado direito da imagem) a medida/métrica QUANTIDADE-FUNICIONARIO (lado esquerdo da imagem). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW. *

Sua resposta

02. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 1, como a inclusão de informações ou sua organização?

Sua resposta

VOLTAR

PRÓXIMA

Página 4 de 7

Avaliação dos conceitos representados no DW

Por favor, responda as questões apresentadas abaixo sobre o Recorte 2:

Recorte 2 de modelo conceitual



03. O Recorte 2 mostra as dimensões disponíveis para se analisar (quadrados com nome iniciando com "dm-") a medida/métrica VALOR-SALARIO-PARTICIPACAO (quadrados com nome iniciando com "mea-"). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW. *

Sua resposta

04. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 2, como a inclusão de informações ou sua organização?

Sua resposta

VOLTAR

PRÓXIMA

Página 5 de 7

Avaliação dos conceitos representados no DW

Fatos são as tabelas que armazenam os indicadores para análise. Medidas (ou métricas) são armazenadas no DW como colunas das tabelas de fato.

Com base nestas definições, por favor responda as questões apresentadas abaixo sobre o Recorte 3:

Recorte 3 de modelo conceitual



05. O Recorte 3 mostra as medidas/métricas (quadrados com nome iniciando com "mea-") contidas no fato FAT_FUNCI (quadrados com nome iniciando com "ft-"). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW. *

Sua resposta

06. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 3, como a inclusão de informações ou sua organização?

Sua resposta

VOLTAR

PRÓXIMA

Página 6 de 7

Avaliação dos conceitos representados no DW

Por favor, responda as questões apresentadas abaixo sobre o Recorte 4:

Recorte 4 de modelo conceitual



07. O Recorte 4 mostra as dimensões disponíveis (quadrados com nome iniciando com 'dm-') inter-relacionadas com os fatos disponíveis (quadrados com nome iniciando com 'ft-'). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW. *

Sua resposta

08. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 4, como a inclusão de informações ou sua organização?

Sua resposta

VOLTAR

ENVIAR

Página 7 de 7

ANEXO IV – Respostas subjetivas do Questionário 2

As informações presentes no Recorte 1 foram suficientes para responder a questão acima? De qual informação sentiu falta?

(6 respostas)

As informações foram suficientes. Não senti falta de informações.

Sim, foram suficientes.

Senti falta de um nome mais semântico para as tabelas como um prefixo, por exemplo (LU_/DIM_ para Look up/Dimensão). Faltou também a representação das métricas das tabelas fato.

Sim

Foram suficientes. Não senti falta de nenhuma informação.

Sim foi suficiente

Comente sobre a dificuldade de responder à questão acima, a estratégia para chegar à resposta ou qualquer outro assunto que achar pertinente.

(5 respostas)

Não encontrei dificuldade em chegar à resposta, apesar de entender que posso ter errado alguma alternativa. A estratégia foi avaliar os relacionamentos e as informações contidas em cada entidade/tabela.

Sugiro apresentar as tabelas fato no centro do diagrama e as tabelas de dimensão na parte periférica do modelo para melhor compreensão ou adotar cores diferenciadas para cada conceito de modelo dimensional, se possível

A falta de uma melhor semântica nos nomes dificultou um pouco. Eu observei os relacionamentos entre as tabelas para identificar o tipo. As tabelas que recebem a cardinalidade muitos são as fato.

Observei os relacionamentos entre as tabelas e seus nomes e atributos.

Não houve qualquer dificuldade, mas o nome das entidades fatos/agregações poderiam ser melhores

As informações presentes no Recorte 2 foram suficientes para responder às questões acima? De qual informação sentiu falta?

(6 respostas)

As informações foram suficientes.

As informações foram suficientes.

Sim

Sim

Para interpretar as informações exige um conhecimento de modelagem dimensional.

Senti dificuldades para definir os níveis da dimensão.

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

(5 respostas)

Encontrei dificuldade em responder as questões por não entender claramente o conceito de níveis de dimensão.

Para achar a resposta, a estratégia é separar as informações que são únicas de cada funcionário daquelas que podem se referir a um conjunto de funcionários.

Para sanar as dúvidas acerca dos níveis da dimensão observei as FKs e a relação entre elas. Observei que não havia um nível hierárquico entre as FKs, mas que elas representavam um segundo nível em direções diferentes em relação a dimensão Funcionário.

Analisei quais itens representam a dimensão de forma única e quais não. Avaliei que os demais itens não poderiam ser agrupados em uma hierarquia.

o atributo "data de nascimento" pode ser confundido com as chaves de acesso da dimensão "dia", porque conceitualmente também é um dia, mas fisicamente não se trata de uma chave estrangeira para a dimensão de dia, é apenas um atributo

As informações presentes no Recorte 3 foram suficientes para responder às questões acima? De qual informação sentiu falta?

(6 respostas)

Sim

Sim

As informações foram suficientes.

Sim.

Foram suficientes.

Identifiquei 3 operações possíveis de roll up, dia -> mês, dia -> ano e mês -> ano, porém não sei ao certo se esta última deveria ser considerada, mas acredito que essa dúvida tem mais a ver com o enunciado da questão (6) do que com o modelo em si. Para mim, o modelo está bastante claro.

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

(5 respostas)

Achei estas questões mais simples pois trata-se de um conceito de fácil compreensão que é a hierarquia entre dia-mês-ano.

Estar atento ao Roll Up de dia para ano. A estratégia seria observar a representação de relacionamentos entre as tabelas.

Nesta questão não encontrei dificuldades.

Pela figura avalio que possa ser feito rol up de dia para mês, de mês para ano e de dia para ano.

Comentado no item anterior

As informações presentes no Recorte 4 foram suficientes para responder às questões acima? De qual informação sentiu falta?

(6 respostas)

As informações foram suficientes.

Sim, porém é necessário conhecimento de modelagem dimensional.

Sim. O fato de haverem métricas repetidas nas duas tabelas infere que uma pode ser agregada da outra. Poderia haver uma referência disso na nomenclatura das tabelas.

Sim

Foram suficientes.

Sim, mas senti falta do tipo de dado de cada coluna

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

(5 respostas)

Não houve dificuldades.

Não houve dificuldade. A estratégia adotada foi verificar nas colunas que não forem PK ou FK aquelas que têm características de indicadores, isto é, passível de análise, agregação, etc.

Senti dificuldade para avaliar se o campo IND_RECIBTO_APOSE, é uma métrica ou não. Como ela não era uma PK ou FK, considerei a mesma uma métrica.

Não tenho certeza se os atributos "NUM_DIA_" são medidas.

Tive duas dificuldades (dúvidas):

- 1) A coluna "IND_RECIBTO_APOSE" parece ser um indicador booleano que pode ou não ser numérico. Se for numérico, ele pode ser considerado como uma quantidade, caso contrário não. Como existe a dúvida eu não contabilizei essa coluna nas respostas acima.
- 2) A tabela "REG_ULT_MES_CARGA" parece ser uma tabela de controle de carga dos temas presentes no contexto analítico em questão, que pode ser confundida com uma dimensão. Não possui nenhuma característica que possa classificá-la como fato/agregação.

As informações presentes no Recorte 5 foram suficientes para responder às questões acima? De qual informação sentiu falta?

(6 respostas)

As informações foram suficientes.

Necessário conhecimento de modelagem dimensional. Facilitaria responder a informação de quantidade de dimensões e hierarquias se fosse um modelo snowflake.

Sim.

Sim

Foram suficientes.

Tive 3 tipos de dificuldade:

1) Fiquei em dúvida se deveria considerar "Ano Anterior" representada pela coluna "NUM_SEQ_ANO_ANT" como uma dimensão específica;

2) Da mesa forma, fiquei em dúvida se deveria considerar a dimensão "Dia" representada pela colunas "NUM_SEQ_DIA_APOSE", "NUM_SEQ_DIA_EXONDO" e "NUM_SEQ_DIA_FALECI".

Em ambos os casos considerei todas elas como dimensões válidas pois está claro que a informação existe, acredito que o problema foi de entendimento do enunciado da questão.

3) A nomenclatura utilizada nas chaves, ora como número sequencial "NUM_SEQ" (chave burra ou sorrogate key), ora como código "COD_" pode causar dúvidas. No caso da entidade "CACAP_CIVIL", existem as duas colunas, mas como o sufixo de ambas é o mesmo, considerei a coluna "COD_CAPAC_CIVIL" apenas como um atributo, provavelmente significando o código operacional dessa entidade, mas não há garantias sobre isso, se o sufixo não fosse o mesmo, eu a consideraria como uma possível chave para outra entidade e portanto, uma possível hierarquia, mesmo não existindo indicação física para isso, ou seja, uma FK declarada.

Senti falta das entidades "FAIXA ETÁRIA", "ESTADO", "SEXO" e "DIA" no modelo em questão.

Comente sobre a dificuldade de responder às questões acima, a estratégia para chegar às respostas ou qualquer outro assunto que achar pertinente.

(4 respostas)

Não houve dificuldade.

A estratégia adotada é observar as FKs existentes nas tabelas de dimensão para obter a quantidade de dimensões e hierarquias.

As informações prestadas anteriormente ajudaram na avaliação desta resposta.

Como já comentei, minha dificuldade maior foi no momento da contagem das dimensões e hierarquias, especificamente no caso das dimensões de tempo "ANO ANTERIOR" e "DIA", mas como também já disse, está claro que essas informações existem e possibilitam análises, inclusive através de agregações.

ANEXO V – Respostas subjetivas do Questionário 3

Qual sua opinião sobre a utilidade de uma representação visual que descrevesse as informações implementadas no DW alinhadas com conceitos de BI? Ex.: Demonstrasse as dimensões e as medidas (métricas) disponíveis, as análises possíveis de se realizar.

(3 respostas)

As rápidas mudanças que afetam as estratégias das corporações, somadas às grandes quantidades de informações produzidas em diversos relatórios, criaram a necessidade de sintetizar esse conjunto enorme de dados em informações de rápida leitura e fácil compreensão. Desta forma, o BI é um instrumento de fundamental importância pois tem o objetivo de suprir parte dessa necessidade e, nesse sentido, a representação visual é uma das formas mais otimizadas para traduzir os dados em valores utilizados nas decisões estratégicas, pois permite uma percepção direta, objetiva, rápida, eficiente e simples das informações. Para tanto, os dados precisam ser trabalhados adequadamente e esse é uma dos grandes desafios a ser enfrentado em um projeto de BI.

Acho que atuaria como catalisador do aprendizado sobre o negócio que está modelado no BI

Seria de grande valia pois permitiria ao usuário explorar melhor a ferramenta.

Existiria vantagem entre uma representação gráfica e uma textual? (3 respostas)

A representação gráfica tem a vantagem de sintetizar e facilitar a percepção das informações e hoje é praticamente impossível conceber um trabalho baseado em conceitos de BI sem esse tipo de abordagem.

Sim

A gráfica pode ser de mais fácil compreensão.

Se essas informações fossem apresentadas com o uso de termos próprios do negócio ao invés dos nomes no DW, seriam mais úteis?

(3 respostas)

Sim. Para o usuário facilitaria o entendimento e tornaria mais intuitivo a aplicação dos dados contemplados no DW.

Naturalmente, quanto mais próximo o BI estiver da linguagem do negócio, mais à vontade e seguros atuarão os usuários finais

Sim, pois facilitaria a aplicabilidade da ferramenta.

01. O Recorte 1 mostra as dimensões disponíveis para se analisar (lado direito da imagem) a medida/métrica QUANTIDADE-FUNCIONARIO (lado esquerdo da imagem). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW.

(3 respostas)

Permite ao usuário identificar as relações com mais facilidade evitando construir pesquisas cujos resultados sejam inconsistentes.

Muito útil durante a fase de aprendizagem do negócio e do ambiente do BI.

É bastante útil pois evita a utilização de uma métrica de forma ineficaz, levando o usuário a uma pesquisa inconsistente.

02. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 1, como a inclusão de informações ou sua organização?

(3 respostas)

As relações das tabelas poderiam ser representadas de forma mais intuitiva como já existem em alguns aplicativos disponíveis no mercado.

Acho que uma navegação pela ferramenta é indispensável para emitir uma opinião mais colaborativa

Não. A representação está satisfatória.

03. O Recorte 2 mostra as dimensões disponíveis para se analisar (quadrados com nome iniciando com "dm-") a medida/métrica VALOR-SALARIO-PARTICIPACAO (quadrados com nome iniciando com "mea-"). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW.

(3 respostas)

[Útil para entender as várias relações existentes entre as variáveis.

Especificamente na visualização pelo celular que estou tendo, a figura apresenta-se muito confusa, comprometendo qualquer visualização lógica.

Bastante útil pois permite ao usuário visualizar as combinações possíveis na construção das pesquisas.

04. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 2, como a inclusão de informações ou sua organização?

(3 respostas)

Os dados poderiam ser estruturados de acordo com algum critério a ser definido pelo usuário, pois existe uma sobreposição de informações na figura.

Categorizar os dados criando classes mais abrangentes, o que traria ganho de entendimento.

Talvez não esteja no exemplo por falta de espaço. Mas caso não esteja previsto, sugiro que seja colocada uma legenda para caracterizar as linhas tracejadas.

05. O Recorte 3 mostra as medidas/métricas (quadrados com nome iniciando com "mea-") contidas no fato FAT_FUNCI (quadrados com nome iniciando com "ft-"). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW.

(3 respostas)

Facilita o entendimento das inter-relações das medidas/métricas e conseqüentemente, a análise dos resultados obtidos.

A ordenação alfabética no recorte 1 facilita a identificação dos itens por usuários mais experientes.

É importante pois permite ao usuário conhecer melhor a estrutura da ferramenta, ou seja, verificar o que está por traz da interface do BI.

06. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 3, como a inclusão de informações ou sua organização?

(3 respostas)

Possibilitar ao usuário estruturar as informações de acordo com suas necessidades.

Ordenação dos campos, com sugestão de ordem alfabética.

Não, está satisfatório.

07. O Recorte 4 mostra as dimensões disponíveis (quadrados com nome iniciando com "dm-") inter-relacionadas com os fatos disponíveis (quadrados com nome iniciando com "ft-"). Descreva sua opinião em relação à utilidade deste tipo de visualização para a tarefa de análise dos dados do DW.

(3 respostas)

Da mesma forma que os itens anteriores, o Recorte permite um entendimento melhor e mais rápido das inter-relações, mas se oferecesse uma interface mais intuitiva ao usuário facilitaria sua utilização.

A apresentação das setas auxilia o entendimento, porém passa uma sensação de insegurança quanto ao correto sentido das mesmas.

Importante para conexão correta entre as dimensões e fatos existentes.

08. Você teria alguma crítica ou sugestão sobre a representação de dados apresentada no Recorte 4, como a inclusão de informações ou sua organização?

(3 respostas)

Informar mais detalhes, como, por exemplo, qual dado que é hierarquicamente superior, relação de dependência dos dados, o domínio.

Não.

Não, está satisfatória.