

UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Yuri Matelli Calazans Luz

**UM DATA WAREHOUSE COMO SUPORTE
ESTATÍSTICO A INFORMAÇÕES TOXICOLÓGICAS**

Florianópolis

2014

Yuri Matelli Calazans Luz

**UM DATA WAREHOUSE COMO SUPORTE
ESTATÍSTICO A INFORMAÇÕES TOXICOLÓGICAS**

Trabalho de Conclusão de Curso submetido ao Curso de Ciências da Computação da Universidade Federal de Santa Catarina para a obtenção do Grau de Bacharel em Ciências da Computação.

Orientador: Prof. Dr. Paulo José de Freitas Filho

Florianópolis

2014

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Luz, Yuri Matelli Calazans

Um data warehouse como suporte estatístico a informações
toxicológicas / Yuri Matelli Calazans Luz ; orientador,
Paulo José de Freitas Filho - Florianópolis, SC, 2014.
45 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro Tecnológico.
Graduação em Ciências da Computação.

Inclui referências

1. Ciências da Computação. 2. Business Intelligence. 3.
Banco de dados. 4. Data Warehouse. 5. Sistema estatístico.
I. Freitas Filho, Paulo José de. II. Universidade Federal
de Santa Catarina. Graduação em Ciências da Computação. III.
Título.

Yuri Matelli Calazans Luz

**UM DATA WAREHOUSE COMO SUPORTE
ESTATÍSTICO A INFORMAÇÕES TOXICOLÓGICAS**

Este Trabalho de Conclusão de Curso foi julgado aprovado para a obtenção do Título de “Bacharel em Ciências da Computação”, e aprovado em sua forma final pelo Curso de Ciências da Computação.

Florianópolis, 10 de Junho de 2014.

Prof. Dr. Vitório Bruno Mazzola

Coordenador do Curso

Banca Examinadora:

Prof. Dr. Paulo José de Freitas Filho

Orientador

Universidade Federal de Santa Catarina

Prof. Dr. Ronaldo dos Santos Mello

Membro da Banca Examinadora

Universidade Federal de Santa Catarina

Prof. Dr. Rivalino Matias Júnior

Membro da Banca Examinadora

Universidade Federal de Uberlândia

Dedico esse trabalho aos meus pais, **Rubens** e **Liliana**, minhas melhores referências! Que sempre, mesmo em momentos difíceis, priorizaram a educação dos seus filhos.

AGRADECIMENTOS

Primeiramente agradeço à professora Marlene Zannin, coordenadora do CIT/SC, pelo apoio para o desenvolvimento desse trabalho; Os treinamentos que fiz, em Chicago e em Curitiba, foram essenciais para o sucesso dessa empreitada.

Agradeço também o coordenador do curso, Vitório Bruno Mazzola, e o secretário, Ivalter Coutinho. Sem dúvida pessoas fundamentais durante minha graduação. Sempre disponíveis para me orientar e resolver problemas, principalmente nas vezes em que precisei contar com prorrogação da minha matrícula na UFSC.

Por fim agradeço aos professores Paulo José de Freitas Filho e Ronaldo dos Santos Mello pelo apoio na finalização desse trabalho em 2014, ao Alexandre Savaris, gerente de projetos do INCoD, pelas dicas durante a estruturação dessa monografia, e ao bolsista do INCoD Maíke de Paula, alocado em 2013 exclusivamente para me ajudar com esse projeto, pelo comprometimento e dedicação.

RESUMO

Uma eficiente análise do grande número de informações envolvidas no fluxo de trabalho de organizações é fundamental. Essa é uma das preocupações do Centro de Informações Toxicológicas de Santa Catarina, responsável pelo desenvolvimento do DATATOX, sistema nacional que permite o registro de ocorrências sobre intoxicações e envenenamentos. Nesse sentido temos nos processos de Business Intelligence propostas que permitem à implementação de sistemas estatísticos. Apoiado em conceitos como software livre e suportado por testes de performance foi então desenvolvido um sistema de análise estatística aplicado aos dados gerados e armazenados pelo sistema DATATOX.

Palavras-chave: Business Intelligence. Data Warehouse. Banco de dados. Sistema estatístico.

ABSTRACT

An efficient analysis of the large amount of information involved in the workflow of organizations is essential. This is a concern of the Toxicological Information Center of Santa Catarina, responsible for the development of DATATOX, national system that allows the recording of poisoning's occurrences. In this sense we have proposed Business Intelligence processes that enable the implementation of statistical systems. Supported by concepts like free software and supported by software performance testing a statistical analysis applied to the data generated and stored by DATATOX system was developed.

Keywords: Business Inteligence. Data Warehouse. Database. Statistical system.

LISTA DE FIGURAS

Figura 1 - Mapa do Brasil, com destaque aos estados com centros de informações toxicológicas	11
Figura 2 - Imagem de divulgação do serviço de plantão 24h do CIT/SC	12
Figura 3 - Exemplo de um modelo de dados entidade-relacionamento	15
Figura 4 - Exemplo de modelo multidimensional, aplicado ao exemplo no item de seção anterior	15
Figura 6 - Relacionamento entre grupos de agentes, tipos de atendimentos e sexos dos pacientes.....	19
Figura 7 - Tela inicial do DATATOX.....	21
Figura 8 - Respectiva aba de cadastro do DATATOX.....	22
Figura 9 - Respectiva aba de cadastro do DATATOX.....	22
Figura 10 - Respectiva aba de cadastro do DATATOX.....	23
Figura 11 - Respectiva aba de cadastro do DATATOX.....	23
Figura 12 - Respectiva aba de cadastro do DATATOX.....	24
Figura 13 - Respectiva aba de cadastro do DATATOX.....	24
Figura 14 - Respectiva aba de cadastro do DATATOX.....	25
Figura 15 - Respectiva aba de cadastro do DATATOX.....	25
Figura 16 - Respectiva aba de cadastro do DATATOX.....	26
Figura 17 - Respectiva aba de cadastro do DATATOX.....	26
Figura 18 - Respectiva aba de cadastro do DATATOX.....	27
Figura 19 - Estrutura multidimensional para manter um histórico do DW	28
Figura 20 - Estrutura multidimensional para implementar uma bridge table.....	28
Figura 21 - Exemplo de consulta executada no Saiku configurado para o DATATOX	30
Figura 22 - Tempo médio de execução de consultas por número de variáveis relacionadas ...	30
Figura 23 - Tempo de resposta de consultas (BD e DW, série azul e vermelha, respectivamente), em milissegundos, observando que entre as observações 41 e 73 o sistema de ETL está migrando dados (as respostas são mais lentas).	31
Figura 24 - Simulação de 20 mil acessos concorrentes ao BD e o DW.....	32
Figura 23 - Distribuição do sexo dos pacientes com fichas registradas no CIT/SC em 2014, considerando 4 distribuições de faixas etárias.....	33
Figura 24 - Gráfico das fichas registradas no CIT/SC em 2014, considerando o meio de atendimento.....	33
Figura 25 - Gráfico da distribuição em turnos de atendimento das fichas registradas no CIT/SC no primeiro semestre de 2014, considerando apenas as segundas-feiras.	34

SUMÁRIO

1	INTRODUÇÃO	11
1.1	CONTEXTUALIZAÇÃO	11
1.2	OBJETIVOS	12
1.2.1	Objetivo Geral.....	12
1.2.2	Objetivos Específicos	12
1.3	JUSTIFICATIVA	13
1.4	ORGANIZAÇÃO DESSE TRABALHO	13
1.5	SOBRE OS TESTES DE PERFORMANCE	13
2	FUNDAMENTAÇÃO TEÓRICA.....	14
2.1	PROJETO DE BANCO DE DADOS.....	14
2.2	MODELAGEM MULTIDIMENSIONAL E DATA WAREHOUSE	15
2.3	OLAP, ON-LINE ANALYTICAL PROCESSING	16
2.4	BI, BUSINESS INTELIGENCE	16
3	ESTADO DA ARTE.....	18
3.1	O CASO DO CENTRO DE INFORMAÇÕES TOXICOLÓGICAS DO RIO GRANDE DO SUL: USO DO COGNOSBI.....	19
4	MATERIAIS E MÉTODOS	21
4.1	APRESENTAÇÃO DO DATATOX.....	21
4.2	DEFINIÇÃO DAS DIMENSÕES DE DADOS.....	21
4.2.1	Cadastro de dados relacionados ao Atendimento.....	22
4.2.2	Cadastro de dados relacionados ao solicitante.....	22
4.2.3	Cadastro de dados relacionados ao paciente.....	23
4.2.4	Cadastro de dados relacionados à exposição	23
4.2.5	Cadastro de dados relacionados ao agente.....	24
4.2.6	Cadastro de dados relacionados à manifestação	24
4.2.7	Cadastro de dados relacionados ao tratamento.....	25
4.2.8	Cadastro de dados relacionados ao exame	25
4.2.9	Cadastro de dados relacionados à história.....	26
4.2.10	Cadastro de dados relacionados ao acompanhamento	26
4.2.11	Cadastro de dados relacionados ao encerramento	27
4.3	MODELAGEM DO DATA WAREHOUSE	27
4.4	SISTEMA DE ETL	29

4.5	ESCOLHA, CONFIGURAÇÃO E CUSTOMIZAÇÃO DA FERRAMENTA.....	29
4.6	DESEMPENHO E TESTES.....	30
4.7	AMBIENTE EXPERIMENTAL.....	32
4.8	RESULTADOS OBTIDOS.....	33
5	CONCLUSÃO.....	35
5.1	SUGESTÕES PARA TRABALHOS FUTUROS.....	35
	REFERÊNCIAS.....	0
	APÊNDICE A - TABELA COM OS TEMPOS OBSERVADOS NA EXECUÇÃO DE RELACIONAMENTOS ENTRE VARIÁVEIS.....	1
	APÊNDICE B - TABELA COM OS TEMPOS OBSERVADOS NA EXECUÇÃO DE CONSULTAS AO BD E AO DW, OBSERVANDO QUE ENTRE AS OBSERVAÇÕES 41 E 73 O SISTEMA DE ETL ESTÁ EM OPERAÇÃO	3
	APÊNDICE C – ARTIGO	6

O Centro de Toxicologia de Santa Catarina (CIT/SC) [...] desenvolve seu trabalho integrando as três grandes áreas de extensão, ensino e pesquisa desde 1983, [...] e atua na pesquisa epidemiológica e clínica, principalmente com as classes de animais peçonhentos, agrotóxicos e medicamentos. Presta também assistência direta ao paciente, ou auxilia o médico através do telefone nas emergências em todo o estado. (CABRAL, 2010).

Figura 2 - Imagem de divulgação do serviço de plantão 24h do CIT/SC



Fonte: Site do Centro de Informações Toxicológicas de Santa Catarina (2014).

É com a informação coletada por esse e pelos demais Centros, originada nas assistências prestadas pelos profissionais, devidamente registrada no DATATOX, que um sistema estatístico teria uma participação fundamental. Surgiu então o anseio por um sistema que reunisse da melhor forma disponível esses dados.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo geral desse trabalho é disponibilizar uma ferramenta para análise estatística dos dados provenientes do DATATOX aos funcionários dos Centros Toxicológicos filiados à ABRACIT, considerando que essa ferramenta deve possibilitar a extração de informações para responder solicitações governamentais e que servirá de apoio à estudantes ou profissionais que sintam necessidade de complementar seus trabalhos acadêmicos.

1.2.2 Objetivos Específicos

- a) Estudar sistemas de código livre disponíveis para análises estatísticas em cubos baseados em consultas OLAP;
- b) Implementar um Data Warehouse como banco de dados do sistema estatístico, evitando assim comprometer o banco de dados transacional;

- c) Desenvolver um sistema de ETL para atualização do Data Warehouse;
- d) Modelar e executar testes de performance para embasar a opção pelo agendamento e execução do sistema de ETL.

1.3 JUSTIFICATIVA

Uma resposta rápida, clara e consistente é fator decisivo para que uma determinada política preventiva cause o impacto esperado. Aqui a participação de um sistema estatístico é fundamental em determinados casos, como deve acontecer no âmbito dos dados registrados pelo DATATOX.

Nesse contexto também é importante ressaltar os possíveis impactos que um sistema estatístico pode trazer a determinada região; Em poder desses dados gestores podem ser responsáveis por decisões que evitem sérias consequências para populações e famílias.

1.4 ORGANIZAÇÃO DESSE TRABALHO

Nos próximos capítulos estão a Fundamentação Teórica, onde são abordadas análises e reflexões que serviram de base para o desenvolvimento desse trabalho, o Estado da Arte, mostrando como essa tecnologia está sendo usada hoje e os Materiais e métodos, apresentando detalhadamente o que foi desenvolvido nesse trabalho.

1.5 SOBRE OS TESTES DE PERFORMANCE

Ao longo desse trabalho serão realizados testes para observar aspectos da implementação, principalmente relacionados à performance dos bancos de dados (transacional, da aplicação, e do Data Warehouse).

A principal métrica será o tempo de resposta, expresso em milissegundos quando não for especificado o contrário. Outra consideração é que esses dados seguem detalhados em anexo (além de apresentados em gráficos em seus respectivos capítulos).

2 FUNDAMENTAÇÃO TEÓRICA

Nos itens seguintes serão abordadas análises e relexões que servem como um suporte teórico à esse trabalho. Do nível mais específico (Projeto de Banco de Dados) ao mais abrangente (Business Intelligence).

2.1 PROJETO DE BANCO DE DADOS

"Um banco de dados é uma coleção de dados relacionados. Com dados, queremos dizer fatos conhecidos que podem ser registrados e possuem significado implícito." (ELMASRI; NAVATHE, 2011)

O armazenamento desses dados normalmente é realizado por um sistema computacional, e sendo feito dessa forma são inúmeras as vantagens: Facilidade de compartilhamento, redução de redundância, controle de consistência, entre outras.

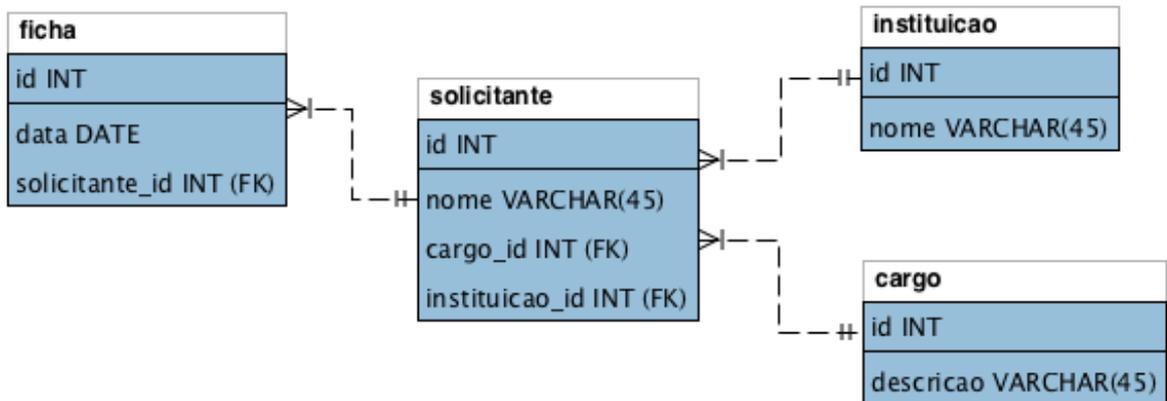
Sistemas transacionais, que podem ser entendidos como aqueles envolvidos nos cotidianos empresariais, tiram proveito dessa situação ao terem um de seus principais alicerces apoiados nas definições e implementações de banco de dados.

Uma das principais características da abordagem de banco de dados é possibilitar a abstração de dados, de modo que diferentes usuários possam percebê-los em seu nível de detalhe preferido. Um modelo de dados - uma coleção de conceitos que podem ser usados para descrever a estrutura de um banco de dados - oferece os meios necessários para alcançar essa abstração. (ELMASRI; NAVATHE, 2011, p. 20).

A grande maioria dos sistemas transacionais desenvolvidos hoje se baseiam em implementações de modelos Entidade Relacionamento, representações de entidades, relacionamentos e atributos para descreverem os dados a serem utilizados.

Porém são processos como a normalização dos dados e a eliminação da redundância que limitam o uso desse modelo, Entidade Relacionamento, em análises estatísticas. Isso porque essas e outras características fazem com que o Modelo Entidade-Relacionamento exija muitas operações de junção para recuperar uma quantidade limitada de informação.

Figura 3 - Exemplo de um modelo de dados entidade-relacionamento



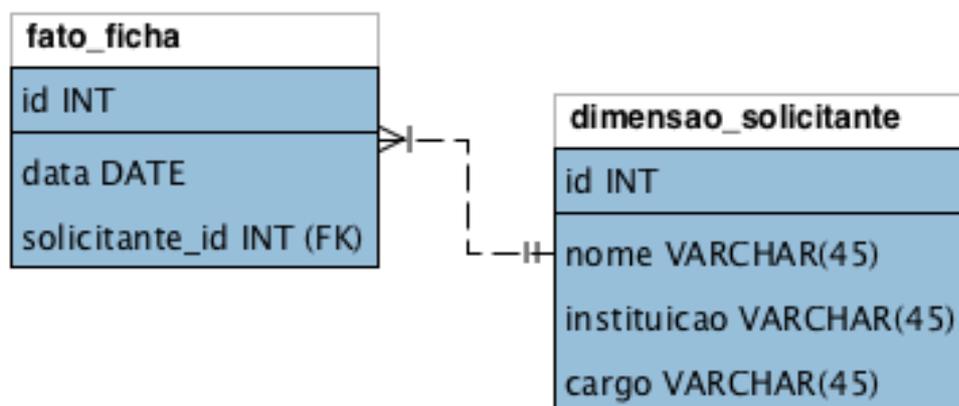
Fonte: Elaborado pelo autor com auxílio da ferramenta Workbench.

2.2 MODELAGEM MULTIDIMENSIONAL E DATA WAREHOUSE

Para responder de forma eficiente demandas para as quais métricas de performance são fundamentais surgiu a modelagem multidimensional, uma alternativa ao modelo Entidade Relacionamento tradicional, que têm os data warehouses como sua principal implementação.

"A dimensional model contains the same information as a normalized model but packages the data in a format whose design goals are user understandability, query performance, and resilience to change." (KIMBALL; ROSS, 2002, p. 12).

Figura 4 - Exemplo de modelo multidimensional, aplicado ao exemplo no item de seção anterior



Fonte: Elaborado pelo autor com auxílio da ferramenta Workbench.

São duas as principais alternativas propostas, Modelo Estrela por Kimball e Modelo Floco de Neve por Inmon. O primeiro, Modelo Estrela, é conhecido por sua simplicidade, já que manter um nível de normalização no data warehouse não é uma de suas preocupações, ao contrário do Modelo Floco de Neve.

Outro conceito importante é o de Datamart, que pode ser visto como um possível subconjunto do Data warehouse. Um data warehouse pode ser formado, por exemplo, de um determinado número de datamarts organizados por assuntos.

Um data warehouse pode ser então definido como um banco de dados de suporte a decisão, ao passo que é contruído de uma maneira que faz com que os tempos de acesso e de resposta a uma série de consultas seja minimizado consideravelmente (em comparação à uma implementação transacional).

2.3 OLAP, ON-LINE ANALYTICAL PROCESSING

É a técnica, ou capacidade, mais popular para explorar data warehouses. Sua importância vai muito além da sua definição, já que ao contrário do que muitos acreditam não se trata de uma ferramenta, mas sim de uma técnica.

Much of the success of OLAP can be attributed to the intuitive approach to data visualization provided by the multidimensional data model. Nowadays, the notions of facts and dimensions have been largely disseminated among database practitioners and researches and have been proved to be useful metaphors to support querying data for decision support. (WREMBEL; KONCILIA, 2007)

OLAP então fornece um poderoso método para empresas acessarem, visualizarem e analisarem seus dados corporativos com alta flexibilidade e performance. Uma consulta que poderia demorar horas em um banco de dados transacional é executada por uma ferramenta OLAP em frações de segundo. É nesse ponto que o tempo de resposta justifica o rótulo de "sistema de tempo-real" atribuído a muitas dessas ferramentas.

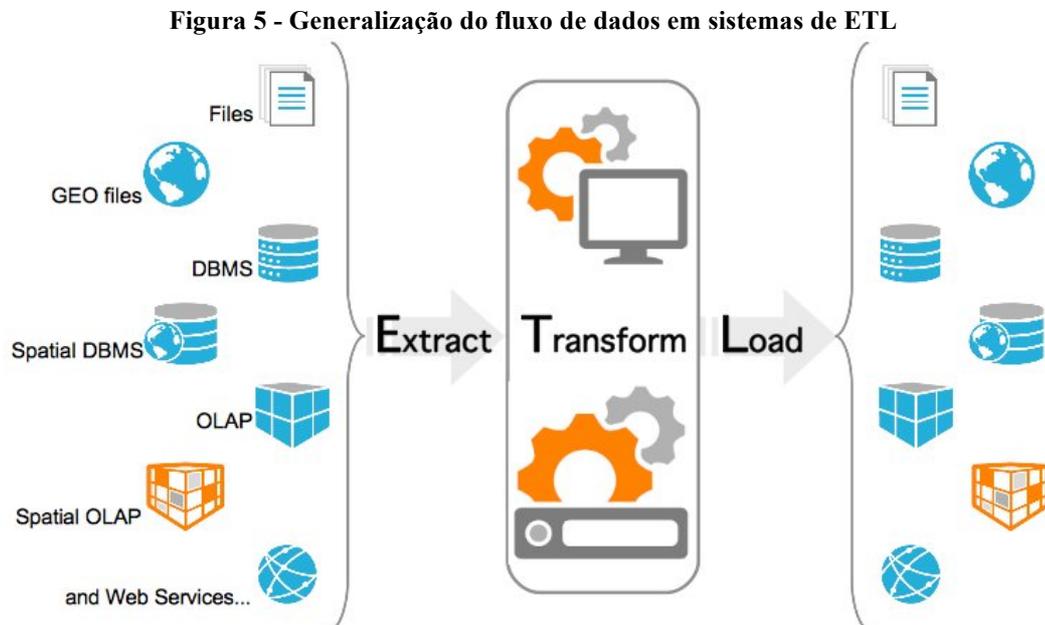
2.4 BI, BUSINESS INTELLIGENCE

Processo também conhecido por Inteligência Empresarial, Business Intelligence refere-se ao uso de técnicas que englobam desde coleta e organização de dados até análise e compartilhamento da informação consolidada.

"Isto significa que é um método que visa ajudar as empresas a tomar as decisões inteligentes, mediante dados e informações recolhidas pelos diversos sistemas de informação". (GCOM, 2013).

Um dos conceitos fundamentais que caracterizam BI é a presença de uma metodologia para ETL (extração, transformação e carga), que através do uso de uma ou mais ferramentas

faz a migração dos dados armazenados em bancos de dados transacionais (entidade-relacionamento) para data warehouses.



Fonte: Kimball e Caserta (2004).

The ETL system makes or breaks the data warehouse. Although building the ETL system is a back room activity that is not very visible to end users, it easily consumes 70 percent of the resources needed for implementation and maintenance of a typical data warehouse. (KIMBALL; CASERTA, 2004, p. XXI).

3 ESTADO DA ARTE

Hoje BI está apoiado sobre a modelagem multidimensional. Mais que uma simples técnica, envolve o que comumente é chamado de ecossistema de inteligência empresarial. No núcleo desse sistema esta a extração, tratamento e carga dos dados, procedimento conhecido pelo acrônimo ETL (Extract, Transform, Load), onde são aplicadas as regras do negócio para a criação de indicadores e variáveis.

Não menos importante, o sistema de armazenamento é fundamental. Modelagem multidimensional nos remete a bancos de dados tradicionais, como os gerenciados pelo MySQL, PostgreSQL, Oracle, mas são estruturas rígidas que muitos estudos vêm contestando. O crescimento e a popularização do termo Big Data deixa esse fenômeno claro, principalmente ao apresentar a alternativa conhecida como NoSQL.

Nesse ponto temos um processo de ETL e um sistema de armazenamento para o Data Warehouse. Precisamos então de uma engine OLAP e de ferramentas, de preferencia visuais e com foco total no usuário final, que permitam uma manipulação com retorno próximo ao real. Isso caracteriza um sistema de Business Intelligence.

Afirmar que existe uma técnica específica que normalize trabalhos de análises de dados toxicológicos seria um erro. Isso porque no Brasil observamos inclusive que é relativamente recente a busca por alternativas ao trabalho manual de cadastros e arquivamentos realizados dentro dos Centros Toxicológicos. Nesse sentido fica claro o pineirismo do CIT/SC ao propor e iniciar o desenvolvimento do sistema transacional DATATOX, e anos mais tarde demonstrar preocupação com a análise desses dados.

No quadro abaixo estão relacionadas características de algumas ferramentas de BI. Essas ferramentas foram selecionadas com base em alguns pré-requisitos, como serem caracterizadas como software livre¹ e terem Java como sua principal linguagem. SAP BusinessObjects, QlikView, IBM Cognos estão entre outras importantes soluções, comerciais, que não estão consideradas nesse estudo.

¹ "Por software livre devemos entender aquele software que respeita a liberdade e senso de comunidade dos usuários. Grosso modo, os usuários possuem a liberdade de executar, copiar, distribuir, estudar, mudar e melhorar o software" (FREE SOFTWARE FOUNDATION, 2014).

Quadro 1 - Características de algumas ferramentas de BI

	Pentaho	SpagoBI	Jaspersoft	OpenReports
Data da análise	11/06/13	11/06/13	27/10/13	18/10/13
Licença	Open source e comercial	Open source	Open source e comercial	Open source e comercial
Duração do projeto	8 anos	8 anos		
Continuidade	Sim	Sim	Sim	Não
Linguagem principal	Java	Java	Java	Java
Controle de versão	CVS	SVN	-	-
Suporte ao Postgres	Nativo (JDBC)	Nativo (JDBC)	-	-

Fonte: Elaborado pelo autor.

3.1 O CASO DO CENTRO DE INFORMAÇÕES TOXICOLÓGICAS DO RIO GRANDE DO SUL: USO DO COGNOSBI

Também um dos precursores quando relacionamos informações toxicológicas com tecnologia da informação, o Centro de Informações Toxicológicas do RS aplicou a solução da IBM, CognosBI, para analisar seus dados. Na imagem abaixo está o resultado do relacionamento entre grupos de agentes, tipos de atendimentos e sexos dos pacientes, em um cubo de dados criado pela IBM para o CIT/RS manipular utilizando essa ferramenta.

Figura 6 - Relacionamento entre grupos de agentes, tipos de atendimentos e sexos dos pacientes

Grupo	Tipo Atendimento		Sexo	
	▪ Humana			
	▪ Total Sexo	▪ Feminino	▪ Ignorado	▪ Masculino
▪ Total Grupo	20.237	10.631	31	9.575
▪ ALIMENTOS	18	10		8
▪ ANIMAIS NÃO PEÇONHENTOS / NÃO VENENOSOS	504	177		327
▪ ANIMAIS PEÇONHENTOS / VENENOSOS	5.548	2.600	4	2.944
▪ ASSOCIAÇÃO DE GRUPOS	480	253		227
▪ DROGAS DE ABUSO	122	36		86
▪ HIGIENE PESSOAL / COSMÉTICOS / PERFUMES	323	162		161
▪ MEDICAMENTOS	6.623	4.286	14	2.323
▪ METAIS	22	7	1	14
▪ NÃO DETERMINADO	1.354	720	1	633
▪ OUTROS	250	109	1	140
▪ PESTICIDAS AGRÍCOLAS	620	188	2	430
▪ PESTICIDAS DE USO DOMÉSTICO	549	272	1	276
▪ PLANTAS / FUNGOS	282	141		141
▪ PRODUTOS DE USO VETERINÁRIO	254	121		133
▪ PRODUTOS QUÍMICOS INDUSTRIAIS	1.185	459	2	724
▪ RATICIDAS	504	270		234
▪ SANEANTES DOMISSANITÁRIOS	1.599	820	5	774

Fonte: Captura de tela no sistema CognosBI.

Uma excelente solução, porém seu caráter comercial limitaria seu uso por Centros espalhados por todo o país, principal objetivo do CIT/SC com o desenvolvimento do

DATATOX. Foram parcerias com órgãos estaduais que permitiram a utilização pelo CIT/RS dessa ferramenta sem custos comerciais.

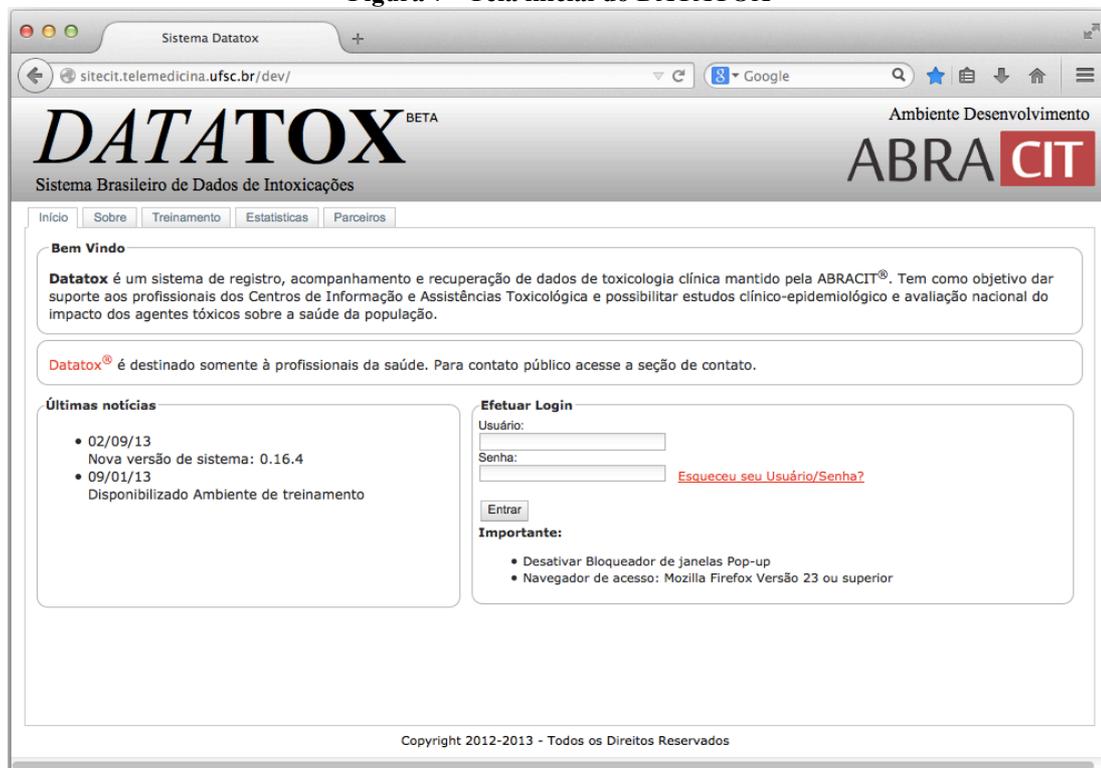
Outra característica limitante é o fato dessa solução se basear em consultas diretas ao banco de dados da aplicação, ou seja, a ferramenta estatística concorre diretamente com a solução transacional. Isso não seria um problema se o DATATOX fosse disponibilizado apenas para SC, mas no momento em que vários Centros tem acesso ao sistema de estatística e ao sistema transacional o número de operações (E/S) tende a aumentar proporcionalmente. A grosso modo, a consequência seria uma "lentidão" em ambos os sistemas.

4 MATERIAIS E MÉTODOS

4.1 APRESENTAÇÃO DO DATATOX

O Sistema de Registro de Intoxicações dos Centros de Informação e Assistência Toxicológica, DATATOX, é um sistema transacional desenvolvido pelo laboratório de Telemedicina da UFSC sob apoio do CIT/SC. Permite o cadastro de várias informações referentes à um determinado incidente toxicológico, como dados dos pacientes, agentes, tratamentos e exames.

Figura 7 - Tela inicial do DATATOX



Fonte: Captura de tela do sistema DATATOX.

4.2 DEFINIÇÃO DAS DIMENSÕES DE DADOS

As dimensões de dados, fundamentais em um sistema estatístico, foram definidas com base nas variáveis geradas pelo DATATOX. Essas variáveis são em maioria qualitativas, ou seja, aceitam um determinado número já existente de alternativas, o que amplia a possibilidade de análises correlatas.

Nas próximas sessões estão capturas de telas de cadastro do DATATOX, e a relação de variáveis consideradas, destacadas em semicírculos de cor vermelha.

4.2.1 Cadastro de dados relacionados ao Atendimento

Figura 8 - Respectiva aba de cadastro do DATATOX

Sistema Datatox

sitecit.telemedicina.ufsc.br/dev/datatox/interno/index

DATATOX^{BETA} Ambiente Desenvolvimento
 ABRA **CIT**

CIT/SC - Centro de Informações Toxicológicas de Santa Catarina

Usuário: yuri Ficha Exposição Humana - Número: 479/2014

Cancelar Ficha Salvar Ficha Resumo Imprimir Fechar Ficha

Atendimento Solicitante Paciente Exposição Agente Manifestação Tratamento Exame História Acompanhamento Encerramento

Atendimento

Meio de Atendimento:
 Telefônico Presencial Outro

Local:

Data/Hora

Data do Atendimento: 02/08/2014 Hora do Atendimento: 16:03

Responsável

Responsável pelo Atendimento:
 yuri

Responsável pelo Preenchimento:
 yuri

Controle do Centro

Código de Registro do Centro

Fonte: Captura de tela do sistema DATATOX

4.2.2 Cadastro de dados relacionados ao solicitante

Figura 9 - Respectiva aba de cadastro do DATATOX

Sistema Datatox

sitecit.telemedicina.ufsc.br/dev/datatox/interno/index

DATATOX^{BETA} Ambiente Desenvolvimento
 ABRA **CIT**

CIT/SC - Centro de Informações Toxicológicas de Santa Catarina

Usuário: yuri Ficha Exposição Humana - Número: 479/2014

Cancelar Ficha Salvar Ficha Resumo Imprimir Fechar Ficha

Atendimento Solicitante Paciente Exposição Agente Manifestação Tratamento Exame História Acompanhamento Encerramento

Dados do Solicitante (?)

Nome:
 Nome Ignorado

Categoria:

País:
 BRASIL

UF:
 SC

Município:
 Município Ignorado

Instituição

UF: SC Município: Instituição:
 Busca: Fonética Pesquisar

Nome:

País:
 BRASIL

UF: SC Município:

Logradouro:

CEP: Bairro: Nº CNES:

Contato

Tipo: Contato:

Tipo Contato Rem.

Não há dados(s) registrado(s)!

Fonte: Captura de tela do sistema DATATOX

4.2.3 Cadastro de dados relacionados ao paciente

Figura 10 - Respectiva aba de cadastro do DATATOX

The screenshot shows the 'Dados do Paciente (?)' registration form in the DATATOX system. The form is divided into several sections: 'Dados do Paciente (?)', 'Residência:', 'Ocupação:', 'Contato', and 'Dados Complementares'. Red boxes highlight the following fields: 'Nome', 'Data de Nascimento', 'Idade', 'Sexo', 'Gestante', 'Peso', 'Raça/Cor', 'Escolaridade', 'País', 'UF', 'Município', 'Ocupação', and the 'Dados Complementares' section. The 'Dados do Paciente (?)' section includes fields for Name, Date of Birth, Age, Sex, Gestational Age, Weight, Race/Color, and Education. The 'Residência:' section includes fields for Country, State, Municipality, Address, City, and ZIP. The 'Ocupação:' section includes a search field and a search button. The 'Contato' section includes fields for Contact Type and Contact Information. The 'Dados Complementares' section includes fields for Mother's Name, CPF, RG, Medical Record, and Other Conventions.

Fonte: Captura de tela do sistema DATATOX

4.2.4 Cadastro de dados relacionados à exposição

Figura 11 - Respectiva aba de cadastro do DATATOX

The screenshot shows the 'Exposição' registration form in the DATATOX system. The form is divided into several sections: 'Exposição', 'Endereço:', 'Via', and 'Circunstância'. Red boxes highlight the following fields: 'Data', 'Hora', 'Tempo decorrido', 'Duração', 'Tipo', 'Local', 'Zona', 'País', 'UF', 'Município', 'Via', and 'Circunstância'. The 'Exposição' section includes fields for Date, Hour, Time elapsed, Duration, Type, Local, and Zone. The 'Endereço:' section includes fields for Country, State, and Municipality. The 'Via' and 'Circunstância' sections each have a dropdown menu and a 'Remover' button.

Fonte: Captura de tela do sistema DATATOX

4.2.5 Cadastro de dados relacionados ao agente

Figura 12 - Respectiva aba de cadastro do DATATOX

The screenshot shows the 'Agentes' tab in the DATATOX system. The interface includes a search bar with a magnifying glass icon and the text 'Pesquisar'. Below the search bar is a table with the following columns: 'Nome', 'Substância/Gênero-Espécie', 'Subclasse', 'Classe', and 'Grupo'. A 'Dados Complementares' field is located below the table, and a 'Remover' button is on the right. The message 'Não há Agente(s) Registrados!' is displayed below the table. The system header includes the DATATOX logo, 'Ambiente Desenvolvimento', and 'ABRA CIT'. The user information is 'Usuário: yuri Ficha Exposição Humana - Número: 479/2014'. The navigation menu includes 'Atendimento', 'Solicitante', 'Paciente', 'Exposição', 'Agente', 'Manifestação', 'Tratamento', 'Exame', 'História', 'Acompanhamento', and 'Encerramento'. The action menu includes 'Cancelar Ficha', 'Salvar Ficha', 'Resumo', 'Imprimir', and 'Fechar Ficha'.

Fonte: Captura de tela do sistema DATATOX

4.2.6 Cadastro de dados relacionados à manifestação

Figura 13 - Respectiva aba de cadastro do DATATOX

The screenshot shows the 'Manifestação' tab in the DATATOX system. The interface includes a 'Classificação Inicial de Gravidade' section with three radio buttons: 'Leve', 'Moderada', and 'Grave'. Below this is the 'Manifestações Clínicas' section, which includes a dropdown menu for 'Manifestações' and a text area for 'Comorbidades'. The system header includes the DATATOX logo, 'Ambiente Desenvolvimento', and 'ABRA CIT'. The user information is 'Usuário: yuri Ficha Exposição Humana - Número: 479/2014'. The navigation menu includes 'Atendimento', 'Solicitante', 'Paciente', 'Exposição', 'Agente', 'Manifestação', 'Tratamento', 'Exame', 'História', 'Acompanhamento', and 'Encerramento'. The action menu includes 'Cancelar Ficha', 'Salvar Ficha', 'Resumo', 'Imprimir', and 'Fechar Ficha'.

Fonte: Captura de tela do sistema DATATOX

4.2.7 Cadastro de dados relacionados ao tratamento

Figura 14 - Respectiva aba de cadastro do DATATOX

The screenshot shows the 'Tratamento' registration page in the DATATOX system. The interface includes a header with the DATATOX logo and 'ABRA CIT' branding. Below the header, there are navigation tabs and a user profile section. The main area is titled 'Tratamento' and contains a form with a 'Tratamentos:' dropdown menu, a 'Categoria:' dropdown menu, and a table with columns for 'Descrição', 'Medida já Tomada', 'Medida Orientada', 'Medida Realizada', 'Ignorado se Realizada', 'Anexo', 'Anexar', and 'Remover Linha'. The table is currently empty, displaying the message 'Não há tratamento(os) registrado(s) !'. Below the table is a section for 'Informações Adicionais Fornecidas'.

Fonte: Captura de tela do sistema DATATOX

4.2.8 Cadastro de dados relacionados ao exame

Figura 15 - Respectiva aba de cadastro do DATATOX

The screenshot shows the 'Exames' registration page in the DATATOX system. The interface includes a header with the DATATOX logo and 'ABRA CIT' branding. Below the header, there are navigation tabs and a user profile section. The main area is titled 'Exames' and contains a form with three dropdown menus: 'Análises Toxicológicas', 'Análises Clínicas', and 'Exames de imagens'. Below these are three dropdown menus for 'Tipo de Exame', 'Data', and 'Hora'. The main area contains a table with columns for 'Descrição', 'Data', 'Hora', 'Complemento', 'Anexos (Resultados)', 'Anexar Resultados', and 'Remover Exame'. The table is currently empty, displaying the message 'Não há exame(s) registrado(s) !'. Below the table is a section for 'Medicamentos Terapêuticos Utilizados'.

Fonte: Captura de tela do sistema DATATOX.

4.2.9 Cadastro de dados relacionados à história

Figura 16 - Respectiva aba de cadastro do DATATOX

Fonte: Captura de tela do sistema DATATOX.

4.2.10 Cadastro de dados relacionados ao acompanhamento

Figura 17 - Respectiva aba de cadastro do DATATOX

Fonte: Captura de tela do sistema DATATOX.

4.2.11 Cadastro de dados relacionados ao encerramento

Figura 18 - Respectiva aba de cadastro do DATATOX

The screenshot shows the DATATOX system interface. At the top, there is a header with the DATATOX logo and 'ABRA CIT' branding. Below the header, there are navigation tabs for various system functions: Atendimento, Solicitante, Paciente, Exposição, Agente, Manifestação, Tratamento, Exame, História, Acompanhamento, and Encerramento. The 'Encerramento' tab is active. The form displays fields for 'Internação' and 'Desfecho', both highlighted with red boxes. Below these fields is a search bar for 'CID - 10' with a 'Pesquisar' button. At the bottom, there is an 'Encerrar Ficha' button.

Fonte: Captura de tela do sistema DATATOX

4.3 MODELAGEM DO DATA WAREHOUSE

Priorizando o grau de liberdade do usuário final na elaboração de suas consultas e considerando a performance de resposta um fator determinante, o data warehouse foi modelado como sendo composto por um único datamart. Isso facilita, entre muitas outras situações, o relacionamento de variáveis dentro do sistema transacional.

O data warehouse foi à princípio implementado em MySQL², mas limitações desse banco de dados obrigaram a equipe de desenvolvimento a substituir esse gerenciador pelo Postgres; Número máximo de índices referenciados por chaves estrangeiras em uma única entidade e controle limitado de transação foram algumas das justificativas.

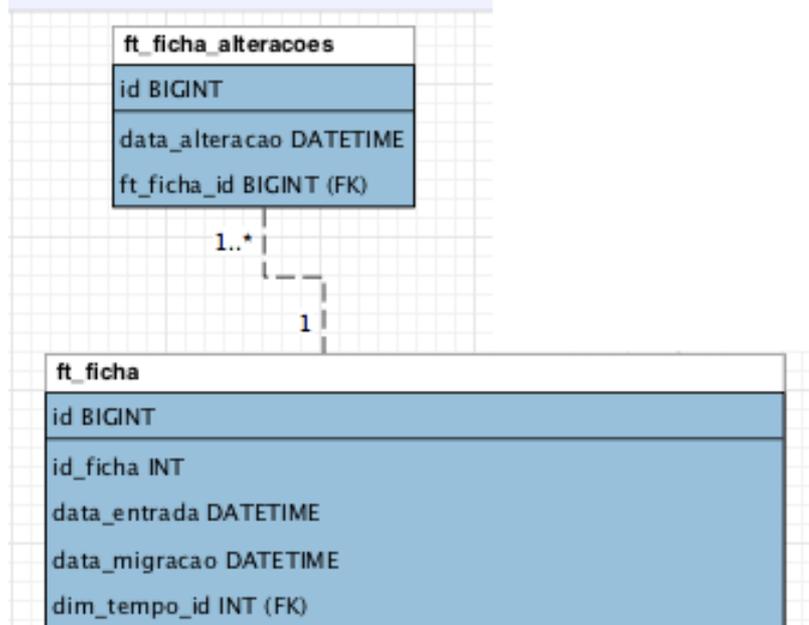
Dentre os principais conceitos usados nessa modelagem podemos destacar dois: O controle de alterações e a modelagem de relacionamentos.

Para manter um histórico consistente optou-se por relacionar à tabela fato (principal tabela em um modelo multidimensional, relacionada a maioria das demais) uma estrutura auxiliar que registrasse o momento em que alterações no data warehouse fossem feitas. Sua

² MySQL e Postgres são sistemas gerenciadores de bancos de dados.

simplifidade favorece sobremaneira a atualização do data warehouse, já que associada à uma chave estrangeira e à uma constraint de operação o registro consolidado pode ser apagado em um único comando.

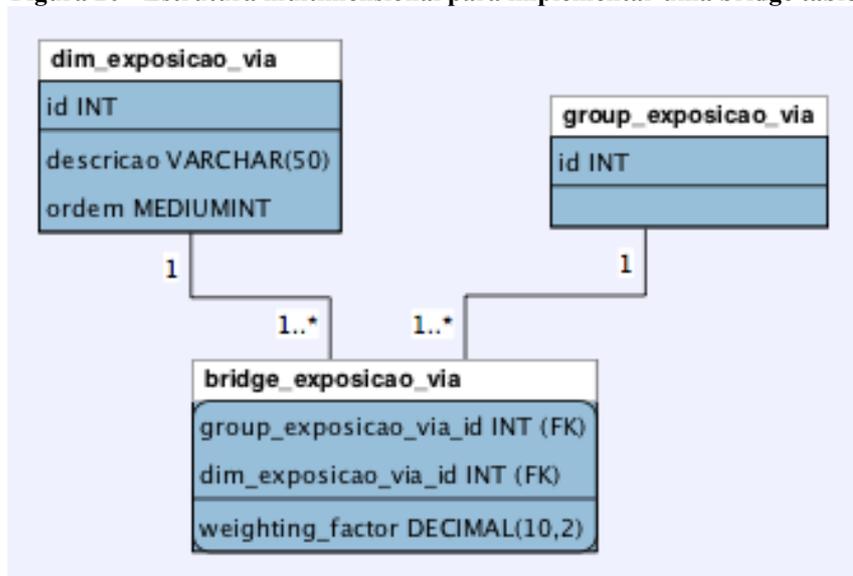
Figura 19 - Estrutura multidimensional para manter um histórico do DW



Fonte: Elaborado pelo autor com auxílio da ferramenta Workbench.

Como alternativa à normalização surge a proposta de se trabalhar com bridge tables, uma forma objetiva de se implementar relacionamentos compostos em tabelas sem que essas precisem recair aos conceitos transacionais de relacionamentos.

Figura 20 - Estrutura multidimensional para implementar uma bridge table.



Fonte: Imagem exportada da ferramenta Workbench.

4.4 SISTEMA DE ETL

Os dados em destaque nas primeiras sessões desse capítulo foram migrados do sistema transacional para o modelo multidimensional por um sistema escrito na linguagem Python. Esse sistema se conecta ao Postgres, faz as respectivas leituras e em uma área reservada da memória realiza a consolidação desses dados, quando necessário, antes de salvar as informações no DW.

Sua execução ocorre uma vez por dia, e esse agendamento é responsabilidade do CRON, que por meio do CRONTAB permite que usuários programem a execução de comandos em ambiente Unix. Os dados consultados no DW refletem então a situação do dia anterior, limitação que garante processamento ao sistema transacional.

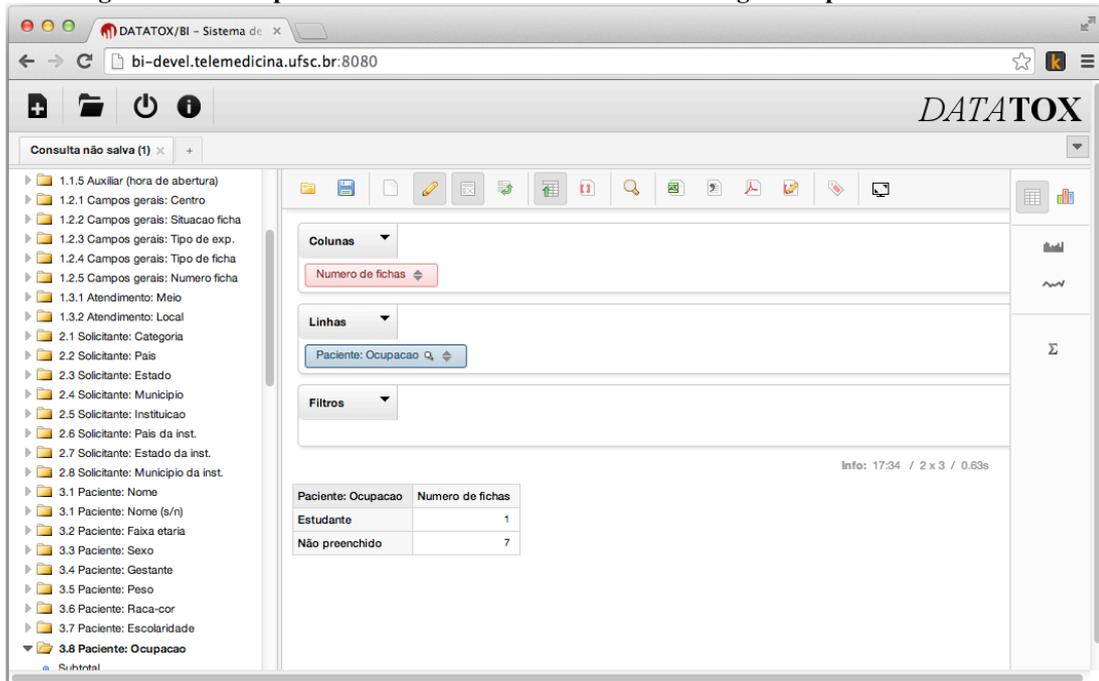
4.5 ESCOLHA, CONFIGURAÇÃO E CUSTOMIZAÇÃO DA FERRAMENTA

O Saiku, ferramenta livre desenvolvida para rodar com o Pentaho, foi selecionada por permitir que de uma forma bastante intuitiva qualquer usuário possa manipular suas consultas. Lembrando que essa precisava ser uma ferramenta compatível com o Mondrian, a base OLAP adotada para realizar consultas no DW.

A customização dessa ferramenta se resumiu em um esforço para concluir uma tradução para o Português Brasileiro e atualizar imagens que exibam o DATATOX como objeto em análise. Nessa etapa também podemos incluir o desenvolvimento de pequenas novas funcionalidades, como o recurso que permite aos administradores salvarem relatórios que serão exibidos como padrão para outros usuários.

Autenticação de usuários em uma base comum ao sistema transacional também foi um requisito no desenvolvimento desse trabalho. Conexões via JDBC, conjunto de classes e interfaces que permitem o envio de instruções SQL para o Postgres, foram suficientes para manter unificado e centralizada essas validações.

Figura 21 - Exemplo de consulta executada no Saiku configurado para o DATATOX

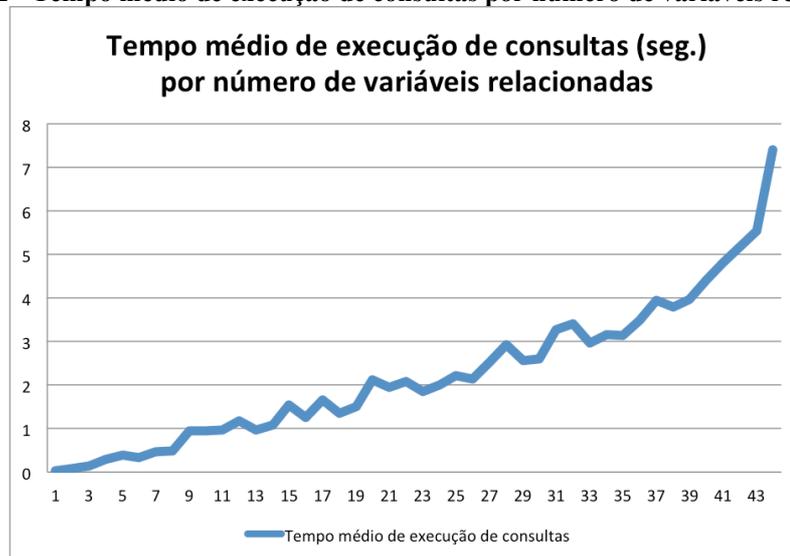


Fonte: Captura de imagem do sistema de estatística do DATATOX.

4.6 DESEMPENHO E TESTES

Os teste seguintes foram realizados com o data warehouse alimentado com 1 mil registros, mas características como modelagem do DW e a distribuição de índices permitem a extrapolação dessas métricas para o mesmo DW com dezenas de milhares de registros. A partir desse ponto é aceita uma degradação, não exponencial, do desempenho das consultas na plataforma OLAP adotada.

Figura 22 - Tempo médio de execução de consultas por número de variáveis relacionadas



Fonte: Elaborado pelo autor com auxílio do Microsoft Excel.

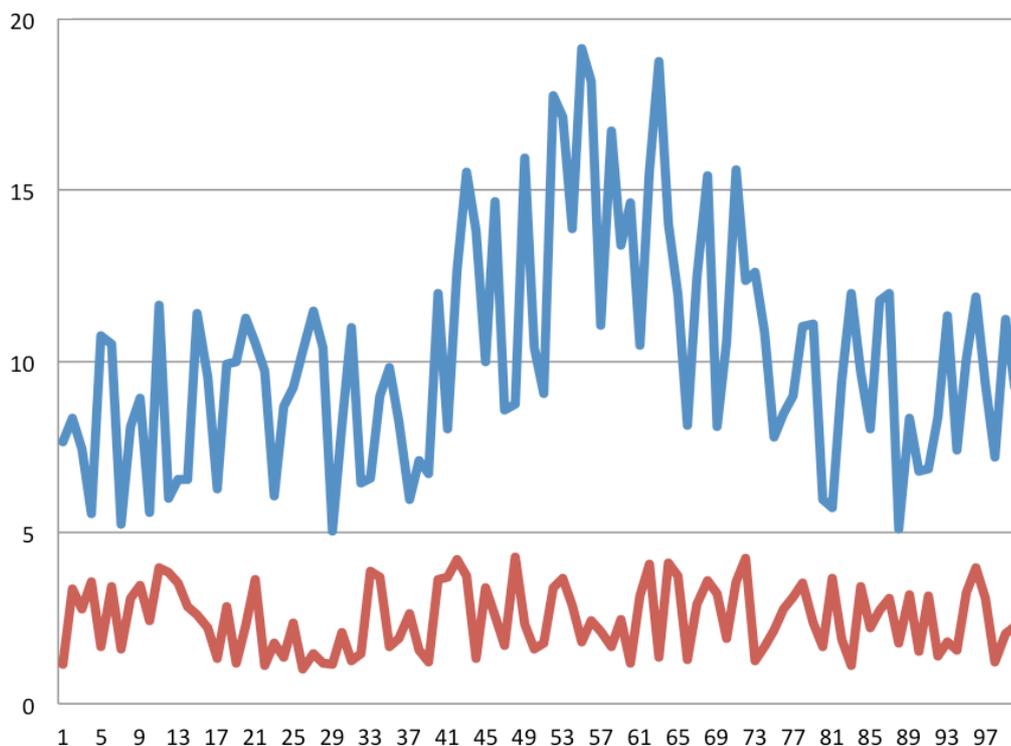
É importante considerar que essas seriam as respostas na primeira execução de um relacionamento único entre as variáveis, já que o Mondrian conta com um sofisticado sistema de cache de consultas, que torna a experiência do usuário final bastante confortável.

Em anexo está uma tabela com as observações que fundamentaram o gráfico acima. Foram 5 execuções dos relacionamentos das variáveis em um ambiente como o descrito no próximo item.

Para viabilizar o desenvolvimento da próxima versão estável, que contará com uma atualização próxima ao tempo real, foram elaborados testes da performance em ambos os bancos de dados, aplicação e Data Warehouse, durante um procedimento de ETL.

Esses testes poderiam inviabilizar essa solução em tempo-real, se nessas condições o tempo de resposta das consultas OLAP no DW sofressem um impacto representativo. Os dados seguem detalhados em anexo, mas depois de extensivos testes podemos concluir que o tempo médio de resposta à consultas no DW aumentou apenas 6% durante a execução do sistema de ETL (considerando que a consulta ao DW envolve o relacionamento de 5 variáveis).

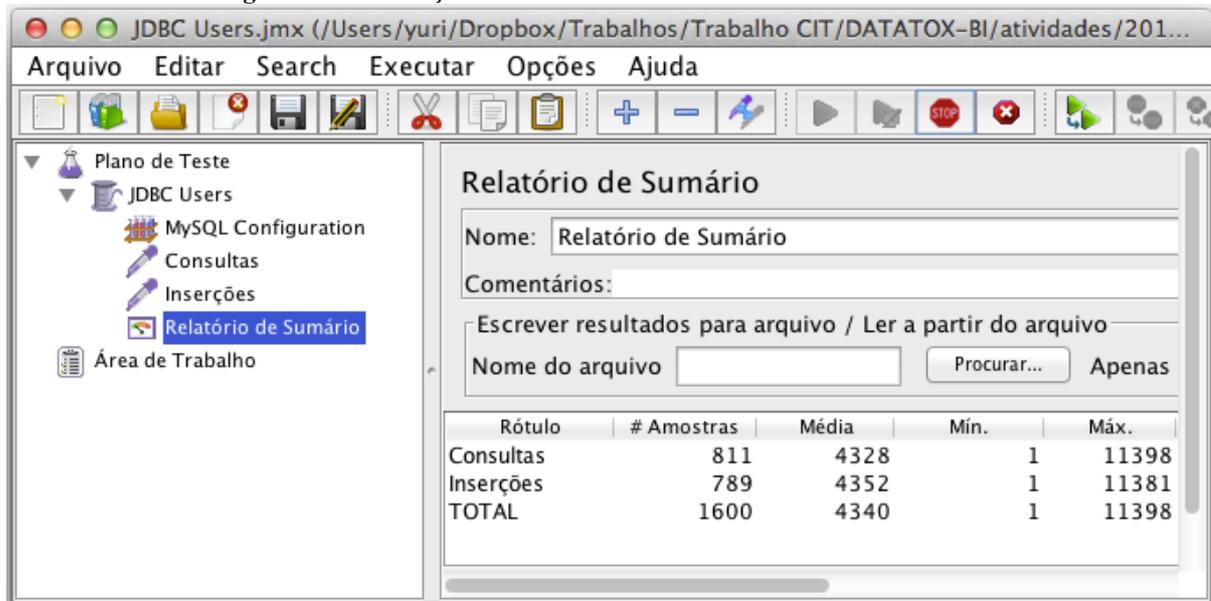
Figura 23 - Tempo de resposta de consultas (BD e DW, série azul e vermelha, respectivamente), em milissegundos, observando que entre as observações 41 e 73 o sistema de ETL está migrando dados (as respostas são mais lentas).



Fonte: Elaborado pelo autor com auxílio do Microsoft Excel.

No teste acima foi utilizado o sistema Apache JMeter para simular a execução de consultas aos bancos de dados por um número estimado de usuários concorrentes.

Figura 24 - Simulação de 20 mil acessos concorrentes ao BD e o DW.



Fonte: Captura de imagem do sistema Jmeter em execução.

4.7 AMBIENTE EXPERIMENTAL

O banco de dados utilizado tanto para a leitura (transacional) quando escrita (data warehouse) foi o Postgres na sua versão 9.1.8, rodando em um servidor virtualizado com capacidade de 100 GB, memória alocada de 4GB (RAM) e acesso a um processador compartilhado Intel Xeon CPU E5-2620 / 2.60GHz. Esse é um fator fundamental a ser considerado, já que a ferramenta OLAP mantém um acesso persistente ao DW mesmo durante a execução de uma consulta.

Os cálculos OLAP foram realizados em uma máquina virtual com capacidade de 183,21GB, memória alocada de 16,43GB e acesso à um processador compartilhado Intel Xeon CPU E5-2620 / 2.00GHz. Apesar do suporte do script ETL à gerencia de threads, por motivos operacionais um único núcleo do processador dessa máquina hóspede foi destinado para a VM.

4.8 RESULTADOS OBTIDOS

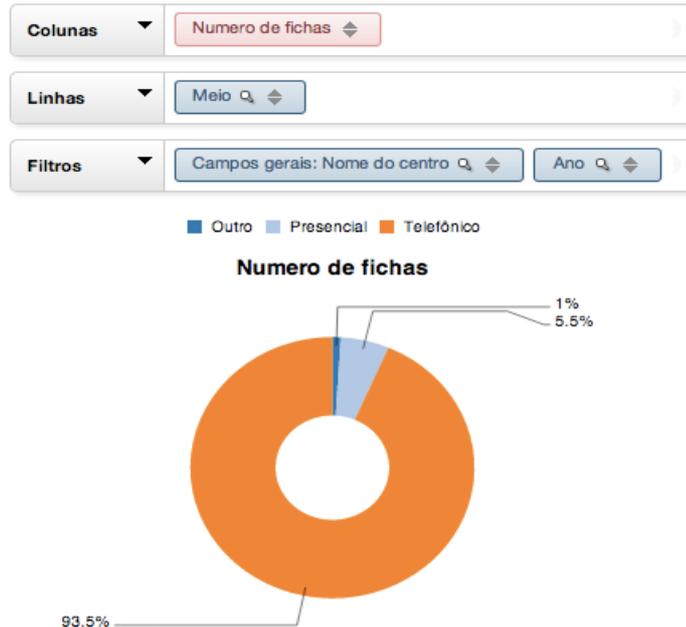
Está disponível o Saiku como sistema estatístico para o usuário final, com o suporte do Postgres como gerenciador do banco de dados e do Mondrian como ferramenta OLAP. A atualização dos dados disponíveis para consulta é feita uma vez ao dia em às 0 hora, momento em que foi observada uma menor utilização do sistema transacional.

Figura 25 - Distribuição do sexo dos pacientes com fichas registradas no CIT/SC em 2014, considerando 4 distribuições de faixas etárias.



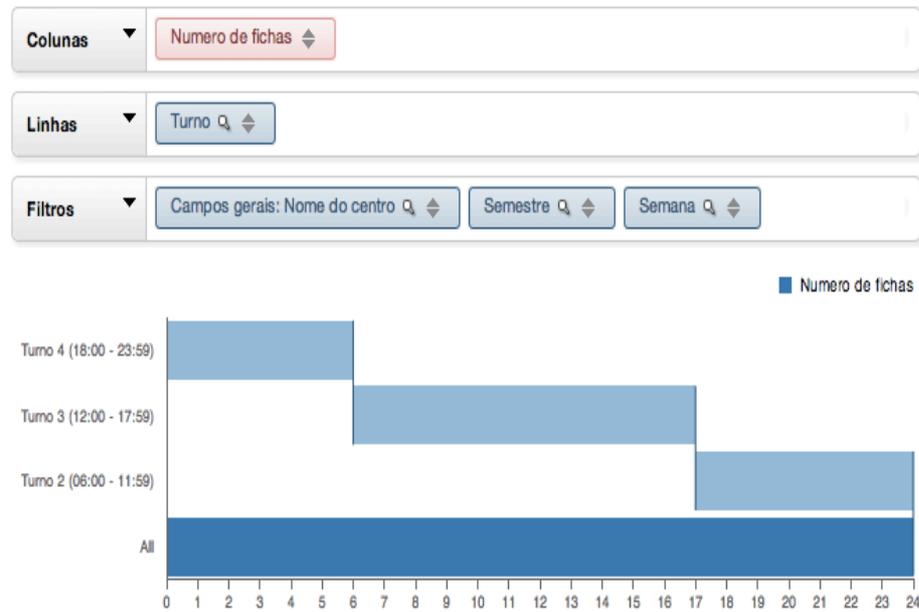
Fonte: Consulta executada no Saiku configurado para o DATATOX.

Figura 26 - Gráfico das fichas registradas no CIT/SC em 2014, considerando o meio de atendimento



Fonte: Consulta executada no Saiku configurado para o DATATOX.

Figura 27 - Gráfico da distribuição em turnos de atendimento das fichas registradas no CIT/SC no primeiro semestre de 2014, considerando apenas as segundas-feiras



Fonte: Consulta executada no Saiku configurado para o DATATOX.

5 CONCLUSÃO

Para que o objetivo geral, disponibilizar uma ferramenta para análise estatística dos dados provenientes do DATATOX, fosse atingido foi feito um completo estudo do estado da arte, que incluiu treinamentos e visitas a Centros Toxicológicos com propostas de análise semelhantes.

No capítulo sobre o estado da arte temos um quadro que lista quatro das muitas ferramentas de BI analisadas ao longo desses últimos meses. A implementação final utilizou componentes da suite Pentaho, mas protótipos foram desenvolvidos com base no SpagoBI, por exemplo.

A facilidade de uso e de configuração da ferramenta Saiku, baseada nas soluções do Pentaho, foi característica marcante para que essa fosse selecionada. Também contribuiu para essa definição a superioridade da engine OLAP Mondrian na manipulação e relacionamento de muitas variáveis, como pode ser constatado na imagem do sub-capítulo Desempenho e Testes.

Conclui-se então que processos de análise que envolvem dados estruturados podem ser enriquecidos com o suporte de Data Warehouses. Nesse contexto a suite de soluções Pentaho e o banco Postgres são ótimas soluções, uma vez que utilizados em conjunto com o Mondrian apresentaram métricas de performance excelentes.

5.1 SUGESTÕES PARA TRABALHOS FUTUROS

Pensando em complementar os processos desenvolvidos nesse trabalho abaixo estão algumas sugestões:

- a) Análise da viabilidade de configurar a ferramenta OLAP para ser usada no contexto da recuperação de informação;
- b) Implementar um controle de inicialização do processo de ETL, que aumente a frequência de atualização do DW;
- c) Aplicar conceitos de programação paralela no sistema de ETL, fazendo que threads controlem a migração de objetos.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE CENTROS DE INFORMAÇÃO E ASSISTÊNCIA TOXICOLÓGICA. c2012. Disponível em: < http://www.abracit.org.br/abracit_site/ >. Acesso em: 30 maio 2014.

CABRAL, Rodrigo Bittencourt. **Concepção, implementação e validação de um enfoque para integração e recuperação de conhecimento distribuído em bases de dados heterogêneas**. 2010. 178 f. Dissertação (Mestrado em Engenharia e Gestão do Conhecimento) - Universidade Federal de Santa Catarina, Florianópolis, 2010. Disponível em: < <http://btd.egc.ufsc.br/wp-content/uploads/2010/12/RodrigoBittencourtCabral.pdf> >. Acesso em: 30 maio 2014.

CENTRO DE INFORMAÇÕES TOXICOLÓGICAS DE SANTA CATARINA. [2014]. Disponível em: < <http://www.cit.sc.gov.br/site/> >. Acesso em: 1 jun. 2014.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de bancos de dados**. 6. ed. São Paulo: Pearson, 2011.

FREE SOFTWARE FOUNDATION. O Sistema Operacional GNU. **O que é o software livre?** 2014. Disponível em: <<http://www.gnu.org/philosophy/free-sw.pt-br.html>>. Acesso em: 1 jun. 2014.

GCOM. c2013. Disponível em: <<http://www.gcom.com.br/Cloud/Home/BI.aspx>>. Acesso em: 5 jun. 2014.

KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit**. 2. ed. New York: Wiley, 2002.

KIMBALL, Ralph; CASERTA, Joe. **The data warehouse toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data**. Indianapolis: Wiley, 2004.

SINTEF. **Big Data, for better or worse: 90% of world's data generated over last two years**. 22 may 2013. Disponível em: <www.sciencedaily.com/releases/2013/05/130522085217.htm>. Acesso em: 10 jun. 2014.

WREMBEL, Robert; KONCILIA, Christian. **Data warehouse and OLAP: concepts, architectures, and solutions**. Hershey; London: IRM Press, 2007.

**APÊNDICE A - TABELA COM OS TEMPOS OBSERVADOS NA EXECUÇÃO DE
RELACIONAMENTOS ENTRE VARIÁVEIS**

Variáveis relacionadas	Tempo médio	Tempo 1	Tempo 2	Tempo 3	Tempo 4	Tempo 5
2	0,032	0,03	0,02	0,05	0,01	0,05
3	0,08	0,07	0,07	0,12	0,1	0,04
4	0,15	0,15	0,18	0,11	0,17	0,14
5	0,288	0,28	0,27	0,28	0,29	0,32
6	0,384	0,38	0,37	0,42	0,41	0,34
7	0,328	0,31	0,3	0,31	0,36	0,36
8	0,466	0,47	0,45	0,5	0,43	0,48
9	0,488	0,51	0,46	0,47	0,5	0,5
10	0,954	0,98	0,94	0,94	0,96	0,95
11	0,958	0,96	1	1,01	0,91	0,91
12	0,97	0,98	0,99	0,93	0,97	0,98
13	1,184	1,19	1,17	1,24	1,17	1,15
14	0,976	0,96	1,01	0,98	0,96	0,97
15	1,074	1,07	1,06	1,04	1,12	1,08
16	1,548	1,55	1,58	1,52	1,5	1,59
17	1,264	1,26	1,29	1,27	1,27	1,23
18	1,668	1,68	1,65	1,63	1,65	1,73
19	1,356	1,36	1,39	1,36	1,34	1,33
20	1,5	1,51	1,46	1,53	1,46	1,54
21	2,126	2,14	2,15	2,11	2,13	2,1
22	1,942	1,93	1,94	1,92	1,98	1,94
23	2,078	2,07	2,03	2,06	2,11	2,12
24	1,844	1,82	1,86	1,84	1,84	1,86
25	2	2	2	2,04	1,96	2
26	2,216	2,23	2,23	2,19	2,24	2,19
27	2,146	2,16	2,17	2,14	2,12	2,14
28	2,528	2,52	2,52	2,56	2,55	2,49
29	2,924	2,92	2,97	2,97	2,87	2,89
30	2,562	2,55	2,58	2,6	2,57	2,51
31	2,602	2,58	2,58	2,63	2,6	2,62
32	3,272	3,26	3,31	3,24	3,25	3,3
33	3,4	3,41	3,37	3,42	3,36	3,44
34	2,97	2,96	3,01	3,01	2,95	2,92
35	3,162	3,17	3,2	3,17	3,13	3,14
36	3,14	3,14	3,15	3,1	3,15	3,16
37	3,48	3,49	3,47	3,52	3,45	3,47
38	3,946	3,98	3,94	3,93	3,93	3,95
39	3,788	3,77	3,81	3,82	3,79	3,75
40	3,962	3,95	3,94	4	3,95	3,97
41	4,412	4,43	4,4	4,39	4,42	4,42

42	4,814	4,8	4,82	4,85	4,82	4,78
43	5,174	5,15	5,2	5,2	5,2	5,12
44	5,55	5,55	5,56	5,51	5,57	5,56
45	7,412	7,4	7,42	7,42	7,37	7,45

Fonte: Elaborado pelo autor com auxílio do Microsoft Excel.

APÊNDICE B - TABELA COM OS TEMPOS OBSERVADOS NA EXECUÇÃO DE CONSULTAS AO BD E AO DW, OBSERVANDO QUE ENTRE AS OBSERVAÇÕES 41 E 73 O SISTEMA DE ETL ESTÁ EM OPERAÇÃO

Observação	Tempo de resposta do BD	Tempo de resposta do DW
1	7,63	1,14
2	8,35	3,36
3	7,43	2,78
4	5,55	3,57
5	10,74	1,68
6	10,51	3,43
7	5,23	1,6
8	8,08	3,08
9	8,92	3,44
10	5,6	2,41
11	11,65	3,97
12	5,99	3,82
13	6,54	3,53
14	6,54	2,84
15	11,38	2,59
16	9,55	2,22
17	6,27	1,31
18	9,91	2,82
19	9,97	1,18
20	11,25	2,27
21	10,53	3,64
22	9,7	1,11
23	6,05	1,78
24	8,67	1,36
25	9,23	2,36
26	10,31	1
27	11,47	1,46
28	10,41	1,18
29	5,04	1,14
30	8,12	2,07
31	10,99	1,27
32	6,43	1,47
33	6,57	3,87
34	8,97	3,68
35	9,8	1,68
36	8,23	1,92
37	5,96	2,64
38	7,08	1,57
39	6,71	1,22

40	11,98	3,64
41	8,03	3,71
42	12,67	4,2
43	15,52	3,73
44	13,79	1,31
45	10	3,38
46	14,67	2,56
47	8,59	1,69
48	8,76	4,28
49	15,92	2,32
50	10,38	1,61
51	9,05	1,78
52	17,74	3,38
53	17,12	3,67
54	13,88	2,88
55	19,11	1,8
56	18,17	2,43
57	11,06	2,13
58	16,73	1,67
59	13,38	2,45
60	14,63	1,18
61	10,48	3,14
62	15,54	4,08
63	18,75	1,36
64	14,01	4,12
65	11,95	3,72
66	8,14	1,3
67	12,46	2,89
68	15,42	3,58
69	8,11	3,23
70	10,58	1,89
71	15,6	3,55
72	12,35	4,25
73	12,61	1,24
74	10,86	1,67
75	7,78	2,15
76	8,48	2,75
77	8,98	3,11
78	11,03	3,51
79	11,08	2,34
80	5,95	1,66
81	5,72	3,65
82	9,42	1,88
83	11,96	1,1
84	9,68	3,41

85	8,02	2,21
86	11,77	2,73
87	11,96	3,07
88	5,1	1,76
89	8,34	3,17
90	6,78	1,52
91	6,86	3,15
92	8,37	1,39
93	11,34	1,81
94	7,39	1,57
95	10,07	3,26
96	11,86	3,96
97	9,32	3,06
98	7,21	1,22
99	11,21	2,05
100	9,25	2,26

APÊNDICE C – ARTIGO

Um Data Warehouse como suporte estatístico a informações toxicológicas

Yuri M. C. Luz¹

¹Departamento de Informática e Estatística – Universidade Federal de Santa Catarina
Campus Universitário, Trindade, Florianópolis - SC, 88040-900

yuri.calazans@gmail.com

Abstract. *An efficient analysis of the large amount of information involved in the workflow of organizations is essential. This is a concern of the Toxicological Information Center of Santa Catarina, responsible for the development of DATATOX, national system that allows the recording of poisoning's occurrences. In this sense we have proposed Business Intelligence processes that enable the implementation of statistical systems. Supported by concepts like free software and supported by software performance testing a statistical analysis applied to the data generated and stored by DATATOX system was developed.*

Resumo. *Uma eficiente análise do grande número de informações envolvidas no fluxo de trabalho de organizações é fundamental. Essa é uma das preocupações do Centro de Informações Toxicológicas de Santa Catarina, responsável pelo desenvolvimento do DATATOX, sistema nacional que permite o registro de ocorrências sobre intoxicações e envenenamentos. Nesse sentido temos nos processos de Business Intelligence propostas que permitem à implementação de sistemas estatísticos. Apoiado em conceitos como software livre e suportado por testes de performance foi então desenvolvido um sistema de análise estatística aplicado aos dados gerados e armazenados pelo sistema DATATOX.*

1. Introdução

São muitos os dados disponíveis hoje, e a tendência é que esse número cresça ainda mais nos próximos anos. Segundo estudos do SINTEF (2013), instituição de pesquisa independente e sem fins lucrativos, "90% dos dados disponíveis hoje foram gerados nos últimos dois anos". As organizações, governamentais ou não, precisam se alinhar a esse novo padrão para não prejudicar a oferta de seus produtos e serviços.

Nesse sentido o Centro de Informações Toxicológicas de Santa Catarina (CIT/SC), filiado a Associação Brasileira de Centros de Informação e Assistência Toxicológica (ABRACIT), se preocupou com o desenvolvimento de uma ferramenta de análise estatística que apoie o DATATOX, Sistema de Registro de Intoxicações dos Centros de Informação e Assistência Toxicológica.

2. Análise tecnológica

Hoje BI está apoiado sobre a modelagem multidimensional. Mais que uma simples técnica, envolve o que comumente é chamado de ecossistema de inteligência empresarial. No núcleo desse sistema esta a extração, tratamento e carga dos dados, procedimento conhecido pelo acrônimo ETL (Extract, Transform, Load), onde são aplicadas as regras do negócio para a criação de indicadores e variáveis.

Não menos importante, o sistema de armazenamento é fundamental. Modelagem multidimensional nos remete a bancos de dados tradicionais, como os gerenciados pelo MySQL, PostgreSQL, Oracle, mas são estruturas rígidas que muitos estudos vêm contestando. O crescimento e a popularização do termo Big Data deixa esse fenômeno claro, principalmente ao apresentar a alternativa conhecida como NoSQL.

Nesse ponto temos um processo de ETL e um sistema de armazenamento para o Data Warehouse. Precisamos então de uma engine OLAP e de ferramentas, de preferencia visuais e com foco total no usuário final, que permitam uma manipulação com retorno próximo ao real. Isso caracteriza um sistema de Business Intelligence.

Afirmar que existe uma técnica específica que normalize trabalhos de análises de dados toxicológicos seria um erro. Isso porque no Brasil observamos inclusive que é relativamente recente a busca por alternativas ao trabalho manual de cadastros e arquivamentos realizados dentro dos Centros Toxicológicos. Nesse sentido fica claro o pineirismo do CIT/SC ao propor e iniciar o desenvolvimento do sistema transacional DATATOX, e anos mais tarde demonstrar preocupação com a análise desses dados.

4. Extração de conhecimento

O Sistema de Registro de Intoxicações dos Centros de Informação e Assistência Toxicológica, DATATOX, é um sistema transacional desenvolvido pelo laboratório de Telemedicina da UFSC sob apoio do CIT/SC. Permite o cadastro de várias informações referentes à um determinado incidente toxicológico, como dados dos pacientes, agentes, tratamentos e exames.

As dimensões de dados, fundamentais em um sistema estatístico, foram definidas com base nas variáveis geradas pelo DATATOX. Essas variáveis são em maioria qualitativas, ou seja, aceitam um determinado número já existente de alternativas, o que amplia a possibilidade de análises correlatas.

O data warehouse foi à principio implementado em MySQL , mas limitações desse banco de dados obrigaram a equipe de desenvolvimento a substituir esse gerenciador pelo Postgres; Número máximo de índices referenciados por chaves estrangeiras em uma única entidade e controle limitado de transação foram algumas das justificativas.

Os dados desse sistema transacional foram migrados para o modelo muldimensional por um sistema escrito na linguagem Python. Esse sistema se conecta ao Postgres, faz as

respectivas leituras e em uma área reservada da memória realiza a consolidação desses dados, quando necessário, antes de salvar as informações no DW.

5. Ferramenta open-source

Para o desenvolvimento desse trabalho foi selecionada a ferramenta Saiku Analitics, sistema de código aberto que recentemente se tornou parte da solução integrada Pentaho.

Reconhecida mundialmente essa suite de ferramentas mantém como sua principal característica a qualidade em que é mantida e público alvo para o qual é divulgada. Abaixo está uma captura da tela desse ferramenta, executando integrada ao Pentaho.

The screenshot shows the Pentaho Saiku Analytics interface. The main window displays a data table with the following structure:

Department	MeasuresLevel	Central	Eastern	Southern	Western
Executive Management	Actual	1,776,282.00	1,507,580.00	1,507,580.00	1,507,580.00
	Budget	2,043,642.00	1,483,508.00	1,523,508.00	1,443,508.00
Finance	Actual	3,106,680.00	3,039,180.00	3,039,180.00	3,039,180.00
	Budget	3,067,361.00	3,010,015.00	3,010,015.00	3,000,015.00
Human Resource	Actual	3,438,863.00	3,212,200.00	3,212,200.00	3,212,200.00
	Budget	3,414,295.00	3,195,682.00	3,195,682.00	3,183,682.00
Marketing & Communication	Actual	3,590,423.00	3,440,110.00	3,440,110.00	3,440,110.00
	Budget	3,582,552.00	3,383,905.00	3,401,905.00	3,401,905.00
Product Development	Actual	2,997,702.00	2,548,800.00	2,548,800.00	2,548,800.00
	Budget	3,159,180.00	2,530,477.00	2,548,477.00	2,548,477.00
Professional Services	Actual	20,068,039.00	18,749,870.00	18,749,870.00	18,749,870.00
	Budget	20,400,000.00	19,100,000.00	18,390,000.00	18,208,206.00

6. Referencias

ASSOCIAÇÃO BRASILEIRA DE CENTROS DE INFORMAÇÃO E ASSISTÊNCIA TOXICOLÓGICA. c2012. Disponível em: http://www.abracit.org.br/abracit_site/. Acesso em: 30 maio 2014.

CABRAL, Rodrigo Bittencourt. Concepção, implementação e validação de um enfoque para integração e recuperação de conhecimento distribuído em bases de dados heterogêneas. 2010. 178 f. Dissertação (Mestrado em Engenharia e Gestão do Conhecimento) - Universidade Federal de Santa Catarina, Florianópolis, 2010. Disponível em: < <http://btd.egc.ufsc.br/wp-content/uploads/2010/12/RodrigoBittencourtCabral.pdf> >. Acesso em: 30 maio 2014.

CENTRO DE INFORMAÇÕES TOXICOLÓGICAS DE SANTA CATARINA. [2014]. Disponível em: < <http://www.cit.sc.gov.br/site/> >. Acesso em: 1 jun. 2014.

ELMASRI, Ramez; NAVATHE, Shamkant B. Sistemas de bancos de dados. 6. ed. São Paulo: Pearson, 2011.

FREE SOFTWARE FOUNDATION. O Sistema Operacional GNU. O que é o software livre? 2014. Disponível em: <<http://www.gnu.org/philosophy/free-sw.pt-br.html>>. Acesso em: 1 jun. 2014.

GCOM. c2013. Disponível em: <<http://www.gcom.com.br/Cloud/Home/BI.aspx>>. Acesso em: 5 jun. 2014.

KIMBALL, Ralph; ROSS, Margy. The data warehouse toolkit. 2. ed. New York: Wiley, 2002.

KIMBALL, Ralph; CASERTA, Joe. The data warehouse toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis: Wiley, 2004.

SINTEF. Big Data, for better or worse: 90% of world's data generated over last two years. 22 may 2013. Disponível em: <www.sciencedaily.com/releases/2013/05/130522085217.htm>. Acesso em: 10 jun. 2014.

WREMBEL, Robert; KONCILIA, Christian. Data warehouse and OLAP: concepts, architectures, and solutions. Hershey; London: IRM Press, 2007.