

**UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE SISTEMAS DE INFORMAÇÃO**

ADAM VARGAS DE LIMA

**DW 2.0: UMA FORMA DE TRATAMENTO PARA DADOS
NÃO ESTRUTURADOS EM ACÓRDÃOS**

**FLORIANÓPOLIS
2011**

ADAM VARGAS DE LIMA

**DW 2.0: UMA FORMA DE TRATAMENTO PARA DADOS
NÃO ESTRUTURADOS EM ACÓRDÃOS**

Trabalho apresentado ao curso de Sistemas de Informação da Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do título de Bacharel em Sistemas da Informação.

Professor: José Leomar Todesco

FLORIANÓPOLIS

2011

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE SISTEMAS DE INFORMAÇÃO

ADAM VARGAS DE LIMA

**DW 2.0: UMA FORMA DE TRATAMENTO PARA DADOS NÃO
ESTRUTURADOS EM ACÓRDÃOS**

Trabalho apresentado ao Curso de Sistemas de Informação da Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do título de Bacharel em Sistemas de Informação.

Prof. Leandro José Komosinski – Coordenador

Prof. José Leomar Todesco – Orientador

Prof. Leonardo de Oliveira Müller – Co-orientador

Prof. Ronaldo dos Santos Mello – Banca

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus e aos meus pais, que com seus esforços me proporcionaram sempre a melhor educação e me deram a oportunidade de estudar em uma universidade de qualidade. Nada disso seria possível sem eles.

Aos amigos, pelo apoio e pelos bons momentos nas horas mais difíceis, quando um tempo longe do trabalho era necessário. Espero que todos saibam da importância que tiveram nesta fase da minha vida.

Ao meu orientador, José Leomar Todesco, e ao meu co-orientador, Leonardo de Oliveira Müller, sempre presentes e dispostos a compartilhar seus conhecimentos.

À Universidade Federal de Santa Catarina pelo meu crescimento nas habilidades técnicas e pessoais, as quais possibilitaram a concretização deste projeto.

E a todos que contribuíram direta ou indiretamente para a realização deste trabalho.

"O mestre disse a um dos seus alunos:
Yu, queres saber em que consiste o conhecimento?
Consiste em ter consciência tanto de conhecer uma coisa
quanto de não a conhecer.
Este é o conhecimento."
Confúcio

RESUMO

Este trabalho de conclusão de curso tem como tema o tratamento e integração de dados não estruturados, assunto que vem ganhando crescente importância e que ainda se mostra bastante desafiador. Este projeto tem como objetivo a obtenção de uma estrutura eficiente para utilização em documentos jurídicos.

O tema será primeiramente apresentado de forma teórica, buscando abranger as definições dos termos, processo de desenvolvimento, origens, vantagens de sua utilização e dificuldades encontradas no seu uso. Na parte teórica do trabalho, também serão apresentados os documentos a serem utilizados, os Acórdãos.

Na parte prática do trabalho será apresentado, em um primeiro momento, o planejamento de todo o trabalho, passando pela análise dos documentos e técnicas para integração dos diferentes tipos de dados. A seguir, serão relatadas a execução dos passos previamente planejados e as dificuldades encontradas. Por fim, serão mostrados os resultados encontrados com a análise dos dados e comentada a possibilidade de trabalhos futuros na área.

Palavras-chave: data warehouse, dados não estruturados, tomada de decisão.

ABSTRACT

This degree dissertation has as its theme the unstructured data treatment and integration, an issue that has become increasingly important and still proves to be quite challenging. This project aims to achieve an efficient structure for use in legal documents.

The theme is first presented in a theoretical way, aiming to cover the definitions of terms, development process, its origins, its advantages and the difficulties encountered in its use. Also, the documents called “Acórdãos” will be presented.

The practical part of the work will present, in a first moment, the project planning, through analysis of documents and techniques for the different data types integration. Next, the execution of the steps previously planned and the difficulties encountered will be reported. Finally, the results will be shown with the data analysis the possibility of other projects in the same area will be commented.

Keywords: data warehouse, unstructured data, decision making.

LISTA DE FIGURAS

Figura 1 - Instâncias da Justiça Estadual	17
Figura 2 - Exemplo de esquema estrela	27
Figura 3 – Exemplo de esquema floco de neve	28
Figura 4 - Ciclo de vida dos dados no DW 2.0	31
Figura 5 - Inclusão dos dados não estruturados	36
Figura 6 - Acesso aos dados não estruturados na base.....	37
Figura 7 - Tratamento dos dados não estruturados.....	39
Figura 8 - Estrutura geral do projeto.....	49
Figura 9 - Exemplo de estrutura no Data Warehouse atual	52
Figura 10 - Tabela fato jurisprudencia, criada no projeto.....	53
Figura 11 - Utilização do plugin desenvolvido.....	56
Figura 12 - Dimensão Dim_lei	59
Figura 13 - ETL da Dimensão Dim_lei	60
Figura 14 - Dimensão Dim_decisao	60
Figura 15 - ETL da Dimensão Dim_decisao	61
Figura 16 - Tabela fato Jurisprudencia.....	62
Figura 17 - Consulta SQL na tabela fato	62
Figura 18 - Exemplos de dados das Dimensões.....	64
Figura 19 - Gráfico do resultado de pedidos de Habeas Corpus para suspeitos de tráfico de drogas	65
Figura 20 - Gráfico de mandado de segurança para professores de licença.....	66
Figura 21 - Gráfico dos processos quanto à mudança na equivalência de ações após venda de companhia	66
Figura 22 - Gráfico relativo à ações de dano moral por inscrição indevida em serviços de proteção ao crédito.	67
Figura 23 - Gráfico dos processos quanto ao ressarcimento de danos em acidentes automotivos.....	68

SUMÁRIO

1. INTRODUÇÃO.....	11
1.1. Objetivos	13
1.1.1. Objetivos Gerais.....	13
1.1.2. Objetivos Específicos.....	13
1.2. Justificativa.....	14
1.3. Metodologia.....	15
1.4. Organização do Trabalho	16
2. FUNDAMENTAÇÃO TEÓRICA	17
2.1. Sistema Jurídico Brasileiro.....	17
2.1.1. O Processo	17
2.1.2. Acórdão.....	18
2.1.3. Jurisprudência	21
2.2. Data Warehouse	24
2.2.1. Dimensões.....	26
2.2.2. Tabelas Fato	27
2.2.3. Esquema estrela	27
2.2.4. Esquema floco de neve	28
2.2.5. Data Mart	28
2.2.6. Top-down.....	29
2.2.7. Bottom-up	29
2.3. DW 2.0	30
2.4. Dados não estruturados	32
2.4.1. Descrição.....	32
2.4.2. Dificuldades	34
2.5. Visão geral dos dados não estruturados no DW 2.0.....	35
2.5.1. Metadados	37
2.5.2. Envolvimento do Usuário	38
2.5.3. Tratamento	39
2.5.4. ETL	40
3. PLANEJAMENTO.....	48
3.1. Proposta do Trabalho.....	48
3.1.1. Análise do Acórdão.....	49
3.1.2. ETL do documento não estruturado.....	50
3.1.3. Integração ao banco de dados estruturado	51
3.1.4. Análises.....	54
4. INTEGRAÇÃO DOS DADOS	54
4.1. Fontes de informação	54
4.1.1. Acórdãos	55
4.1.2. Dados estruturados.....	55
4.2. Integração com o DW.....	55
4.3. Extensão Desenvolvida	56
4.3.1. Tratamento de Acordao.....	56

4.4. Processo de ETL.....	59
4.4.1. Dimensão Lei	59
4.4.2. Dimensão Decisão	60
4.4.3. Jurisprudência	62
5. RESULTADOS	63
6. CONCLUSÕES	69
6.1. Trabalhos Futuros.....	69
7. REFERÊNCIAS BIBLIOGRÁFICAS	70

1. INTRODUÇÃO

O sistema judiciário brasileiro possui hoje um grande banco de dados, o qual conta com o apoio de um sistema de Data Warehouse para geração de estatísticas e relatórios. Este DW é separado por instâncias do judiciário, as quais explicaremos melhor no capítulo sobre o Sistema Judiciário, e por estados da federação. Seu principal objetivo é prover informações sobre os processos.

Este sistema, assim como a quase totalidade dos sistemas de apoio à decisão atuais, foi desenvolvido visando o tratamento e um melhor armazenamento dos dados estruturados, estes obtidos diretamente de um banco de dados relacional. Os dados não estruturados não foram levados em conta neste desenvolvimento inicial.

O motivo mais provável da sua não consideração é de que a ideia da sua utilização é recente, já que estes arquivos não foram previstos nas primeiras versões de Data Warehouse. Seu uso é também consideravelmente mais complexo se comparado aos tradicionais dados estruturados, tanto na sua extração quanto na integração ao DW.

Devido à crescente perda de informação que ocasionava, esta limitação dos conceitos iniciais do DW foi sendo cada vez mais notada.

Então, quando da criação do conceito de DW 2.0 por Bill Inmom, uma das principais mudanças em relação ao antigo modelo foi justamente uma maior atenção aos dados não estruturados.

Neste trabalho, utilizaremos estes dados para criarmos estatísticas sobre um conceito jurídico que vem ganhando cada vez mais força no nosso Sistema Judiciário: a jurisprudência.

A jurisprudência, definida por Miguel Reale, como “a sucessão harmônica de decisões dos tribunais”, ou seja, interpretações e decisões semelhantes para casos semelhantes, nunca foi considerada nesse sistema.

Enquanto nos países da Common Law, direito desenvolvido a partir das decisões dos tribunais, utilizado nos países de colonização inglesa, a jurisprudência pode definir o rumo de um processo. No nosso país este conceito foi, por muito tempo, relegado a um papel secundário. Isso se deve muito em função da herança do Direito Romanístico, linha do Direito esta que considera a Lei escrita como único meio pelo qual um processo deve ser julgado.

Ultimamente seu uso vem sendo cada vez mais discutido. Muitos estudiosos consideram a jurisprudência como uma possível solução para a tão criticada lentidão do nosso sistema judiciário. Apontam, para isso, o fato de que se o magistrado pudesse ter como ponto de partida um caso semelhante, ou, até mesmo, usar como decisão a repetição de um julgamento de um processo anterior, a conclusão de um processo poderia ser muito mais rápida.

Sobre este contraponto à herança do Direito Romano, Luís Felipe de Freitas Kietzmann comenta que “mesmo em face do nosso sistema jurídico essencialmente legislativo, não se pode ignorar a dinâmica da sociedade moderna, que torna necessário ao aplicador das leis, frequentemente, recorrer às demais fontes do direito”.

Mesmo entre alguns críticos do uso da jurisprudência como caráter normativo, ou seja, como poder de decisão, há a consideração de que pode ser usada senão como fonte principal, como auxílio à decisão.

Atualmente, a única ferramenta de auxílio à jurisprudência é uma busca textual simples, em que se buscam palavras semelhantes em processos. Neste cenário não é possível a geração de estatísticas e também não existe a integração com o Data Warehouse. Diante destas mudanças que vem sendo discutidas, é importante que o sistema de apoio à decisão também forneça as informações necessárias para lidar com estas novas ideias e possíveis alterações na ciência jurídica.

Considerando estas dificuldades, este trabalho busca, através das técnicas de DW 2.0, um tratamento e uma integração eficiente para estes dados, de forma que possam ser criadas estatísticas sobre a jurisprudência que supram a atual carência.

1.1. Objetivos

1.1.1. Objetivos Gerais

O objetivo geral deste trabalho é a extração dos dados não estruturados e sua integração com os dados estruturados do atual sistema, buscando a geração de estatísticas sobre a jurisprudência.

1.1.2. Objetivos Específicos

- Apresentação das vantagens da extração da informação de dados não estruturados.
- Compreensão das necessidades de informação a serem extraídas dos acórdãos que possam ajudar no processo da obtenção de informações sobre o uso da jurisprudência.
- Projeto para criação física das novas tabelas no DW já existente, buscando a correta integração dos dados não estruturados.

- Implementação de extensões à uma ferramenta de integração de dados para suprir as necessidades do tratamento dos dados não estruturados em questão.
- Criação de consultas e estatísticas visando o apoio à jurisprudência.

1.2. Justificativa

Durante muito tempo os desenvolvedores de sistemas de Data Warehouse mantiveram o foco apenas nos dados estruturados do banco de dados transacional. Os dados não estruturados recebiam pouca ou nenhuma atenção, fato este ocorrido em grande parte devido à falta de percepção sobre a quantidade de informação que poderia ser obtida utilizando-os.

A importância destes dados foi notada diretamente na prática das empresas. Muito se perdia sem o armazenamento de dados encontrados em arquivos como e-mails e notas de texto.

O que já vinha sendo notado na prática, Bill Inmon expôs no livro DW 2.0 TM - Architecture for the Next Generation of Data Warehousing.

Mesmo com sua importância sendo reavaliada, o conceito de tratamento e integração de dados não estruturados atualmente ainda possui raríssimos exemplos de sua utilização no mercado.

Este novo paradigma vem ao encontro das necessidades observadas neste trabalho, quem tem como objetivo um eficiente tratamento dos acórdãos utilizando técnicas e conceitos criados por Inmon.

Ao extrairmos e estruturarmos a informação da lei usada nas decisões judiciais, esta hoje só encontrada nos arquivos não estruturados, poderemos descobrir padrões extremamente úteis, que poderão auxiliar no que tange ao uso desta fonte do direito. Visto que, em poder da lei utilizada, tornaremos possível a análise do que motivou o magistrado a tomar determinada decisão, a partir disto poderemos inferir se estas mesmas motivações foram usadas em outros casos parecidos.

Teremos então dados suficientes para a geração de estatísticas sobre o uso da jurisprudência.

Acreditamos que este projeto seja de grande valia para o aumento do uso e maior entendimento dos benefícios que podem ser adquiridos com estes dados.

1.3. Metodologia

Este trabalho dividiu-se em 5 partes:

- O primeiro passo foi a análise dos dados não estruturados, os acórdãos, e a busca pelo tipo de informação útil para a jurisprudência. Nesta etapa foram analisadas também as dificuldades no tratamento e na integração destas informações no sistema atual.
- Após a análise de quais dados do acórdão poderiam ser interessantes, foi desenvolvido uma extensão de um software para integração de dados chamado Kettle. Esta extensão, denominada plugin, tem como função efetuar a extração das informações apontadas como necessárias para o trabalho.
- De modo adjacente ao desenvolvimento do plugin, foi feito o planejamento para integração do DW existente com os dados não estruturados a serem carregados a partir deste trabalho e então criada a estrutura necessária no banco de dados.
- Uma vez pronto o desenvolvimento e a estrutura das tabelas para o processo de carga dos dados, foi executada a extração, o tratamento e a carga das informações, utilizando o Kettle.
- Depois de feita a inclusão dos dados no DW, foram possibilitados os relatórios, estatísticas e consultas tendo como assunto a jurisprudência.

1.4. Organização do Trabalho

Neste primeiro capítulo foram apresentados o motivo do trabalho, os objetivos, a sua justificativa e a metodologia utilizada. O segundo capítulo traz a literatura envolvendo os conceitos básicos de Data Warehouse, informações sobre dados não estruturados, conceitos novos presentes no DW 2.0 e a utilização de dados nesta nova proposta. No terceiro capítulo é mostrado detalhadamente o planejamento do trabalho prático, compostos pela Análise dos Acórdãos, desenvolvimento a extensão ao software de integração de dados e a criação das dimensões e tabelas fato necessárias.

O capítulo seguinte mostra os resultados da execução do planejamento do capítulo anterior e seus resultados. Por último, no quinto capítulo, são relatadas as conclusões deste projeto e os trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Sistema Jurídico Brasileiro

Para entendermos o motivo deste trabalho precisamos conhecer alguns pontos que envolvem um processo e suas diferentes fases, além da estrutura do acórdão e do conceito de jurisprudência.

2.1.1. O Processo

Primeiramente, vamos conhecer de uma forma extremamente simplificada, apenas para entendimento do projeto, como é o andamento de um processo.

Na figura 1, vemos o percurso que um processo pode percorrer no Sistema Judiciário Estadual.

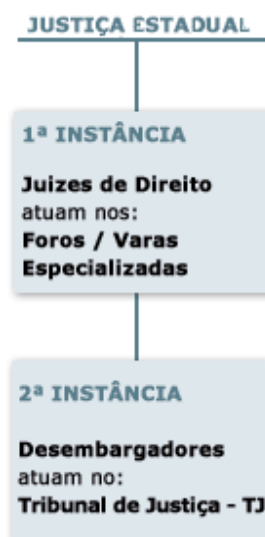


Figura 1 - Instâncias da Justiça Estadual

Quando da entrada de um processo, primeiramente ele ingressa na primeira instância, ou primeiro grau. Em seguida é feita a distribuição do processo para algum magistrado da vara em questão, e então, após o julgamento, o juiz responsável por este processo dá a sua sentença.

Se uma das partes não concorda com a sentença dada em primeira instância, recorre à justiça de segundo grau para obter uma nova decisão.

Nesta segunda instância, o processo será analisado por um conjunto de magistrados, denominados desembargadores. O documento desta nova decisão é chamado de acórdão. Este documento é o foco do nosso trabalho.

Conheceremos melhor então a estrutura e as informações do acórdão.

2.1.2. Acórdão

Conforme o Vocabulário Jurídico Conciso (DE PLÁCIDO E SILVA), um acórdão é “a resolução ou decisão tomada coletivamente pelos tribunais”. O que o diferencia da sentença, é, então, o seu caráter de decisão por um órgão colegiado, ao contrário da sentença que é emanada de um órgão monocrático.

Simplificando, o acórdão é a decisão tomada por um grupo de desembargadores, os juízes dos Tribunais de Justiça dos Estados, proferido por um tribunal de justiça de segundo grau, que pode manter ou alterar a sentença do juiz de primeiro grau.

A estrutura geral de um acórdão e suas definições conforme De Plácido e Silva costuma ser:

❖ Dados gerais;

Informações iniciais do processo em questão, divididas em:

- Número de identificação da apelação.

A apelação designa um dos recursos de que se pode utilizar a pessoa prejudicada pela sentença, a fim de que, subindo a ação à superior instância, e,

conhecendo esta de seu mérito, pronuncie uma nova sentença, confirmando ou modificando, a que se proferiu na jurisdição de grau inferior.

- Apelante.

Diz-se da parte litigante, ou do terceiro prejudicado, que intentou o recurso de apelação sobre sentença, que lhe tenha causado gravame ou provocado prejuízo.

- Apelado.

Termo que designa a pessoa, que teve sentença favorável, de que se apelou.

Exemplo no acórdão:

Apelação nº 990.09.000054-6

Apelante: Paulo Rauen

Apelado: Joana Luz

❖ Ementa.

O resumo que se faz dos princípios expostos em uma sentença ou em um acórdão, ou o resumo do que se contém numa norma.

No acórdão pode ser exemplificado por:

APELAÇÃO CÍVEL. AÇÃO CIVIL PÚBLICA. ATO DE IMPROBIDADE ADMINISTRATIVA. COLOCAÇÃO DE SERVIDORES EM ÓCIO PROPOSITAL. PERDA PATRIMONIAL E VIOLAÇÃO AOS PRINCÍPIOS DA ADMINISTRAÇÃO.

Constitui ato de improbidade administrativa a conduta do agente que mantém servidores em estado de ócio, remunerados pelos cofres públicos, sem que houvesse, em contrapartida, a devida prestação do serviço. A sanção de ressarcimento deve limitar-se ao período em que os servidores mantiveram-se em inatividade. Recurso conhecido. Preliminar rejeitada. Recurso parcialmente provido.

❖ Relatório.

Designa a exposição ou a narração, escrita ou verbal, acerca de um fato ou de vários fatos, com a discriminação de todos os seus aspectos ou elementos.

Por exemplo, o início de um relatório no acórdão pode ser:

Trata-se de recurso de apelação interposto contra a sentença de fls.223/227, que julgou procedente o pedido da ação civil pública por ato de improbidade administrativa e condenou Paulo Rauen às sanções previstas no art.12, incisos II e III, da Lei n.º 8.429/92, bem como ao pagamento de honorários advocatícios no importe de R\$5.000,00 (cinco mil reais).

❖ Voto.

Manifestação da vontade, ou a opinião manifestada, pelo membro de uma corporação, ou de uma assembleia, acerca de certos fatos e mediante sistema, ou forma, preestabelecida.

No acórdão, o magistrado explica o motivo de seu voto, baseando-se na constituição, para justificar sua decisão. Pode se iniciar desta maneira:

Cuidam os autos de ação civil pública por ato de improbidade administrativa, em que o Ministério Público do Estado de Minas Gerais pretendeu imputar ao ex-Prefeito Municipal de Paraobepa (José Antônio de Matos), ao ex-Secretário Municipal de Administração e Recursos Humanos (Fábio Botelho Porto) e ao ex-Secretário Municipal de Obras (Roberto Carlos Franco), as condutas tipificadas nos artigos 10, caput, e 11, inciso I, da Lei n.º 8.429/92, em razão da suposta colocação de servidores municipais em situação de ócio, motivada por razões de perseguição política.

O pedido foi julgado procedente e os apelantes condenados às penas de ressarcimento ao erário, perda da função pública, suspensão dos direitos políticos, pagamento de multa civil e proibição de contratação com o Poder Público, sendo esta a razão do inconformismo recursal.

Com efeito, verifica-se que os recorrentes José Antônio de Matos e Fábio Botelho Porto, por meio do Decreto n.º 149/2002 (fls.66/69) e das Portarias n.º 465/2002 (fls.62/63) e 470/2002 (fls.64/65), colocaram em disponibilidade os servidores Pedro Luis, Wagner Antônio e Wanderley Gomes, os quais,

posteriormente, reingressaram no serviço público por força de decisão judicial transitada em julgado (fls.55/59).

❖ **Decisão.**

Solução que é dada a uma questão ou controvérsia, pondo fim a ela, por meio de sentença, despacho ou interlocutória, e criando uma nova composição entre as partes contendoras ou litigantes. É, assim, o resultado de um pleito, quando é tida num sentido mais estrito, ou a mera deliberação a respeito de um ato ou de qualquer pedido que se faz no processo, numa acepção mais ampla.

Por exemplo:

Isso posto, DOU PARCIAL PROVIMENTO ao recurso de apelação para 1) ajustar a condenação de ressarcimento ao período em que os servidores mantiveram-se em inatividade; 2) decotar da sentença as sanções de proibição de contratação com o poder público e perda da função pública; 3) reduzir os honorários advocatícios para R\$1.500,00 (um mil e quinhentos reais). Custas ex lege.

2.1.3. Jurisprudência

A definição de jurisprudência varia muito na linguagem técnica jurídica conforme o autor.

Segundo Diniz (1993, p.290), jurisprudência é o conjunto de decisões uniformes e constantes dos tribunais, resultante da aplicação de normas a casos semelhantes, constituindo uma norma geral aplicável a todas as hipóteses similares e idênticas. É o conjunto de normas emanadas dos juízes em sua atividade jurisdicional.

De Plácido e Silva (2009) a descrevem modernamente como (...) o hábito de interpretar e aplicar a lei aos fatos concretos, para que, assim, se decidam as causas. Desse modo, a jurisprudência não se forma isoladamente, isto é, pelas

decisões isoladas. É necessário que se firme por sucessivas e uniformes decisões, constituindo-se em fonte criadora do Direito e produzindo um verdadeiro jus novum. É necessário que, pelo hábito, a interpretação e explicação das leis a venham formar.

Para Miguel Reale (1998, p. 167), ela significa "a forma de revelação do direito que se processa através do exercício da jurisdição, em virtude de uma sucessão harmônica de decisões dos tribunais".

Interpretando estes autores, concluímos então, que o conceito de jurisprudência seria a consideração de mesmas ou semelhantes interpretações à casos análogos, utilizando como base para isso o registro de outras decisões anteriores e adaptando-os conforme as particularidades do processo em questão.

Com o maior uso dessa técnica, vantagens poderiam ser adquiridas. De acordo com Ernesto Junior Silveira Netto, "A jurisprudência evitaria que uma questão doutrinária ficasse eternamente aberta e desse margem a novas demandas, portanto diminuiria os litígios, reduziria os inconvenientes das incertezas do Direito, porque faria saber qual seria o resultado das controvérsias. Uma das maiores causas de queixas do sistemas judiciário é a lentidão, a jurisprudência viria em socorro desta demanda, possibilitando uma maior rapidez nas decisões uma vez que fornece subsídios valiosos ao magistrado."

A principal vantagem do seu uso seria a segunda citada, a de uma maior rapidez no nosso sistema judiciário, como também afirma Luís Felipe de Freitas Kietzmann "Tem-se reconhecido cada vez mais a importância da jurisprudência no ordenamento jurídico pátrio, mormente quando se discute alternativas para desembaraçar o Poder Judiciário."

A jurisprudência, conforme as tradições e origens do Direito, é mais utilizada em países que adotam o sistema da *Common Law*, como os Estados Unidos, onde pode ser vital em uma ação judicial. No Brasil, atualmente, pesquisas jurisprudenciais não são consideradas de suma importância. Esta característica vem da herança romanística do Direito brasileiro, o qual defende que o único Direito é a lei escrita, sendo função do Poder Judiciário apenas a sua interpretação.

A mudança deste cenário vem sendo discutido cada vez mais. O professor doutor Fredie Didier Jr., defensor da importância da Jurisprudência como Fonte do

Direito Processual, fala sobre a precedência nos julgamentos: “Toda decisão judicial traz, em seu bojo, no mínimo duas espécies normativas ou normas jurídicas. A primeira é a norma jurídica individualizada que é a que consta do dispositivo das decisões judiciais e que regula o caso concreto submetido à apreciação do Poder Judiciário.”

Prova dessa mudança de paradigma é que mesmo nos países herdeiros do Direito romano, segundo Kietzmann “não mais prevalece a redução do Direito à lei, muito embora subsista a primazia desta sobre todas as outras fontes. Assim, mesmo em face do nosso sistema jurídico essencialmente legislativo, não se pode ignorar a dinâmica da sociedade moderna, que torna necessário ao aplicador das leis, frequentemente, recorrer às demais fontes do direito.”

O uso da jurisprudência como fonte de direito é um ponto que gera muita discussão entre diversos autores de diferentes correntes. Do mesmo modo que já mostramos posições a favor, existem discordâncias.

Greco Filho (1996, p.369) discorre que é possível dividir em duas correntes doutrinárias a concepção acerca da jurisprudência enquanto fonte de direito, quais sejam (i) a que reconhece sua função criadora de normas, e (ii) a que entende que a jurisprudência se limita a reconhecer e declarar a vontade concreta da lei.

Citando Gustav Radbruch, para quem "os atos jurídicos e as sentenças realizam o direito, mas não influem em sua existência lógica, podendo influir em sua compreensão histórico-cultural", Greco Filho defende que a posição predominante é a segunda, não admitindo, portanto, sua força normativa.

Dinamarco (2004, p. 102) defende que a jurisprudência não é fonte do direito. Admite, porém, que a jurisprudência pode exercer influência sobre as decisões dos julgadores:

A repetição razoavelmente constante de julgados interpretando o direito positivo de determinado modo (jurisprudência) exerce algum grau de influência sobre os futuros julgadores, mas não expressa o exercício do *poder*, com os predicados de generalidade e abstração inerentes à lei.

Sabe-se, portanto, que a jurisprudência efetivamente atua como referência do Julgador em casos análogos, sobremaneira quando os tribunais superiores já se

pronunciaram uniformemente acerca do tema, representando a jurisprudência, na prática, um poder de ditar a aplicação da lei.

Mesmo entre os estudiosos que não concordam com que seja atribuída no Brasil a mesma importância recebida pela jurisprudência em outros sistemas jurídicos, há a ideia de que “nem por isso é secundária a sua importância” (Reale, 1998, p.167), e que ela pode ser considerada uma ferramenta para “indicar o caminho predominante em que os tribunais entendem de aplicar a lei, suprimindo, inclusive, eventuais lacunas desta última” (Martins, 1994, p.58).

Podemos concluir então, que, apesar da discussão sobre possuir ou não competência normativa, existe a concordância entre diversos autores de que existe a possibilidade de que a análise de decisões anteriores de processos semelhantes possa trazer benefícios à justiça.

Vemos também que esta análise atualmente não é usada em sua totalidade, mas vem sendo cada vez mais reconhecida e defendida por vários estudiosos como uma forma de aperfeiçoamento do nosso sistema judiciário.

2.2. Data Warehouse

Primeiramente, os mecanismos de armazenamento de dados eram simples, e não se preocupavam com nada além do puro armazenamento. Não tinham o intuito de buscar informações ou de apresentar dados de uma maneira que pudessem trazer novos tipos de estratégia às empresas. E de certa forma, isto nem era possível, já que cada dado custava muito quando armazenado em cartões perfurados.

Com a introdução das fitas magnéticas essa primeira barreira foi vencida, o armazenamento passou a ser mais barato. Mas ainda assim a leitura era difícil. Foi então que surgiu o armazenamento em disco, facilitando a última barreira, o acesso aos dados.

A revolução foi grande, poderíamos agora guardar dados de forma relativamente barata e acessá-los, apagá-los ou reescrevê-los quando quiséssemos. O que acarretou no surgimento de uma massa de dados nunca antes vista, e, posteriormente, em um grande problema.

Dados com diferentes arquiteturas, em lugares diferentes, por vezes replicados e sem um padrão definido para o mesmo tipo. Tudo isso dificultava muito a obtenção de qualquer informação estratégica.

Foi nesse ambiente de caos que surgiu a idéia do processo de Data Warehousing, que basicamente era reunir esses dados, de forma que fosse desfeita toda essa bagunça em que haviam se transformado, a fim de gerar um suporte gerencial e estratégico às empresas.

Desde o início, um data warehouse, buscando ser a base de todo o processo informacional, teve definições claras, são elas:

- Orientado à assunto
- Integrado
- Não-volátil
- Histórico
- Uma coleção de dados para suporte à decisão

O DW contém dados históricos integrados e granulares. Este é o seu grande segredo, esta integração possibilita às empresas uma capacidade de enxergar seu ambiente como um todo. Informações vindas de diversos lugares diferentes são apresentadas ao cliente como tendo uma fonte em comum. Essa é a primeira grande vantagem, a possibilidade de enxergar através da organização de um ponto de vista coletivo, sem as divisões técnicas dos sistemas anteriores.

A granularidade variável, podendo-se obter diferentes níveis de detalhamento, é outra grande vantagem de um data warehouse. Ela permite a flexibilidade dos dados. Como os dados são granulares, eles podem ser observados de uma maneira por um grupo e de outra maneira por outro grupo, dependendo de seus interesses. Por exemplo, enquanto para um setor que cuida da estratégia de negócios futuros

de uma empresa é interessante a apresentação da informação de uma maneira, para o setor responsável pela distribuição é preferível, e até necessária, outra forma de ver os dados.

Outra grande vantagem surgida com o conceito de Data Warehouse foi o armazenamento histórico dos dados. Possibilitando uma visão a longo prazo, mais estratégica que a tradicional, que mostra apenas a informação atual. Isto torna possível observar a evolução da empresa.

Mas a implementação de um DW não é fácil. É preciso lidar com diversas dificuldades, como a integração dos dados vindos de diferentes fontes, o volume de dados, que por serem históricos e não-voláteis tendem a crescer exponencialmente e a diferença de desenvolvimento. Enquanto outros sistemas são construídos de uma vez só, um DW é construído em várias partes, tanto por seu tamanho, que tende a ser muito grande, quanto por seus requisitos, que tendem a mudar conforme o andamento do projeto.

Por todas suas características já descritas, o DW acaba sendo um processo caro, complicado e de alto risco. O que faz com que as empresas pensem duas vezes antes de executá-lo.

Alguns conceitos da arquitetura e desenvolvimento de um DW serão explicados a seguir. Primeiramente veremos os conceitos de dimensão e fato, após isso os dois esquemas no qual um Data Warehouse pode ser modelado, o que é um Data Mart e, por último, seus métodos de desenvolvimento.

2.2.1. Dimensões

Uma Dimensão é uma tabela que armazena aspectos textuais de cada elemento parte do processo.

As tabelas dimensionais contêm vários atributos que descrevem em detalhes todas as características que possam definir e serem úteis para futuras pesquisas no Data Warehouse.

Podemos usar como exemplo uma dimensão para os Filmes de uma locadora. Nesta dimensão seriam armazenadas características dos filmes como Nome, Produtora, Atores, etc.

2.2.2. Tabelas Fato

Enquanto as dimensões são tabelas textuais que buscam a descrição dos elementos, as tabelas fato tem como objetivo a análise quantitativa.

Tabelas fato contem chaves estrangeiras para as dimensões e sua função é de medição. Por exemplo: número de locações de determinado gênero de filme.

2.2.3. Esquema estrela

Este tipo de modelagem é bastante simples e basicamente é composto de uma única tabela fato ligada a diversas dimensões.

A simplicidade deste sistema é justamente uma de suas vantagens, já que proporciona facilidades ao usuário, o qual criará consultas com relacionamentos de baixa complexidade.

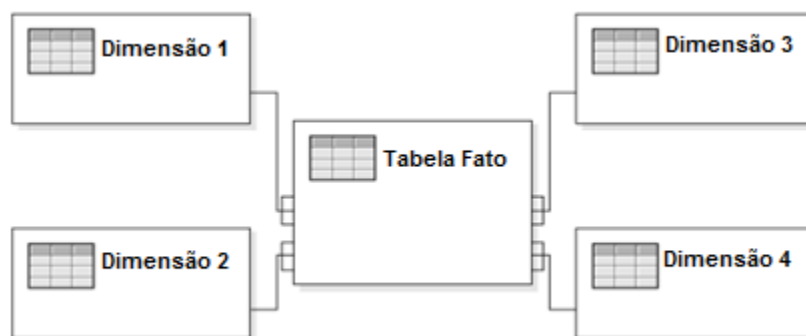


Figura 2 - Exemplo de esquema estrela

2.2.4. Esquema floco de neve

Mais complexo que o esquema estrela, nesta modelagem as dimensões não necessariamente apontam para uma tabela fato.

Contrariamente ao esquema explicado anteriormente, busca a normalização das tabelas, com diversas dimensões podendo ser usadas para um tipo de informação. Para um Produto de uma companhia por exemplo.

Suas consultas geralmente são mais complicadas do que o esquema explicado anteriormente e primam mais pela rapidez na busca dos dados.

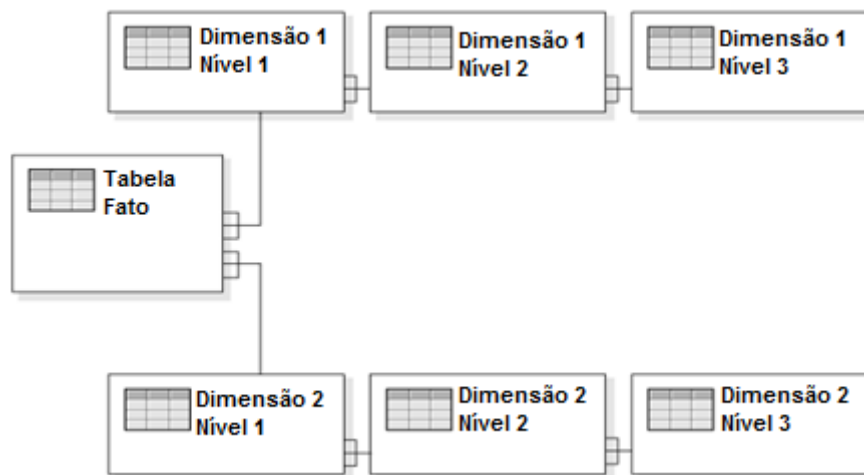


Figura 3 – Exemplo de esquema floco de neve

2.2.5. Data Mart

Data Mart é um subconjunto de um Data Warehouse. São geralmente desenvolvidos a partir das necessidades de uma área da empresa ou do negócio. Como veremos a seguir, podem derivar de um Data Warehouse previamente construído, ou ter como fonte de dados a base transacional e, então, seu conjunto formar o DW.

Por exemplo, em um tribunal de justiça, poderíamos ter um Data Warehouse de todas as informações que envolvem o Processo, e diversos data marts especificando uma informação sobre esta área de negócio, como Tempo de Andamento do Processo, Movimentação do Processo, etc.

2.2.6. Top-down

Este método de desenvolvimento, defendido por Bill Inmon (2002), propõe que seja desenvolvido primeiramente um grande sistema de DW aonde serão concentradas todas as informações da organização relevantes à tomada de decisão independentemente da área. Para uma definição simples, podemos citar Jukic(2006) que cita o Data Warehouse como sendo “uma fonte de dados para os novos Data Marts”. A partir desta fonte de dados central seriam criados os diversos Data Marts necessários. É um modelo consistente quanto à mudanças no negócio. Sua principal desvantagem é o custo para desenvolvimento do grande projeto inicial, tanto na questão de tempo de implementação quanto de pessoas para isto.

2.2.7. Bottom-up

Ao contrário de Inmon, Ralph Kimball (1998) defende que o projeto seja executado a partir de projetos pequenos, os Data Marts, para então, o conjunto destes compor o Data Warehouse da organização.

Neste método de desenvolvimento, os dados para o Data Mart vem diretamente da base transacional, sendo esta a fonte de dados.

Como os Data Marts são independentes, assim que um deles é finalizado a organização já pode usa-lo. Como exemplo, se o Data Mart “Vendas” foi desenvolvido, e o “Marketing” ainda está sendo criado, a empresa já pode criar estatísticas a partir do primeiro.

Esta é uma grande vantagem, fazendo com que a companhia obtenha resultados rapidamente, mesmo que ainda não tenha um Data Warehouse completo.

2.3. DW 2.0

Evolução do DW. É assim que podemos tratar o DW 2.0 descrito por Inmon no livro *DW 2.0 TM - Architecture for the Next Generation of Data Warehousing*. Bill Inmon sendo um dos elaboradores do conceito inicial de Data Warehouse, com o tempo identificou alguns problemas deste primeiro trabalho. Alguns observados, através da experiência após anos de implementação no mercado profissional, outros causados pela falta de registro do conceito.

Para entendermos essa nova abordagem, se torna necessário primeiramente uma visão geral sobre as limitações da primeira ideia de um armazém de dados.

Surgido academicamente nos anos 80 e com forte intuito de ser usado comercialmente, após ser colocado em prática, foram observados novos problemas e requisitos a serem sanados pelos sistemas de apoio à decisão. Como descreve Inmon, “muitas forças moldaram a evolução da arquitetura de informação a este mais alto nível – DW 2.0”.

Entre elas podemos citar a demanda por diferentes tecnologias. A evolução já comentada relacionada ao armazenamento de dados trouxe ao usuário uma forma de interação e uma quantidade de informação cada vez maior. Conforme citado anteriormente, em pouco tempo a computação passou de cartões perfurados a um armazenamento e processamento de dados que tornaram possíveis a criação de telas com uma incrível diversidade de informações, apresentadas das mais diferentes maneiras.

Devido a esta quantidade de informação, passou a ser preocupante e um objeto de estudo o tempo e a facilidade de acesso à elas. Notou-se então que um dado de cinco anos atrás tem um padrão de acesso totalmente diferente de um dado do mês corrente, o que levou a proposta do ciclo de vida de um dado.

De acordo com a proposta de ciclo de vida do dado no DW 2.0, um dado é tratado de diferentes formas conforme sua necessidade de acesso. Necessidade esta que é baseada na “idade” do dado. De acordo com essa teoria, quanto mais antiga a informação, menor é a necessidade de performance na sua busca. São usados então, para dados mais antigos, formas de armazenamento que privilegiem o tamanho da base ao invés da rapidez na sua recuperação. O que faz com que o DW tenha um custo muito menor e uma agilidade muito maior no acesso aos dados.

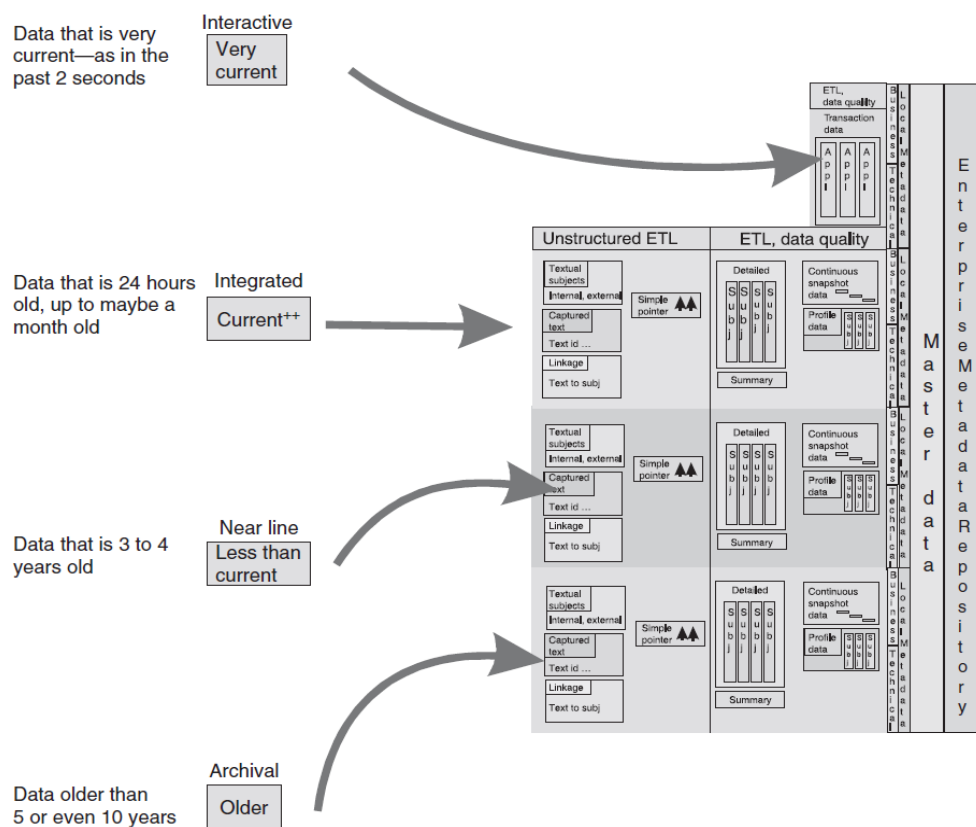


Figura 4 - Ciclo de vida dos dados no DW 2.0

Outro problema surgido com a experiência no mercado foi a necessidade de inclusão de dados não estruturados. Muitas informações podem ser adquiridas de e-

mail, notas, relatórios, etc. Nada disso havia sido considerado na primeira abordagem.

O reconhecimento da importância desses requisitos levou ao DW 2.0, e é a implementação desta última característica o motivo deste projeto.

2.4. Dados não estruturados

2.4.1. Descrição

Hoje é consenso entre os estudiosos de banco de dados a existência de três tipos de dados: estruturados, semi estruturados e não estruturados.

Como o nome já sugere, a grande diferença está em sua forma. Enquanto o primeiro vem sempre em um mesmo formato e layout, e no segundo podemos considerar alguma estrutura, como nos arquivos XML, no terceiro é a organização das informações é até certo ponto imprevisível. Não é seguido um formato, um layout ou mesmo um padrão. Além disso, podem ser totalmente diferentes um do outro. Nessa sua imprevisibilidade reside a dificuldade de tratamento. Apesar destas complicações, a experiência das empresas vem mostrando que vale a pena transcorrer essa barreira.

Desde o desenvolvimento dos primeiros Data Warehouses, as decisões foram sempre baseadas em dados obtidos de tecnologias estruturadas, negligenciando assim as formas de informação não estruturadas, como e-mail, notas e relatórios.

Inmon comenta no artigo Unstructured Applications: Unlocking the Potential, “Cruzando o vazio entre dados não estruturados e dados estruturados, uma nova perspectiva de dados é possível. Fazendo uma ponte entre o vazio entre os dois tipos de ambiente é possível combinar texto e dados numéricos. Esta habilidade possibilita novos tipos de sistemas totalmente novos a serem construídos”.

Um exemplo do que Inmon afirma é a rapidez de informação que um e-mail pode trazer. Este tipo de dado nos mostra informação imediata de um contato com o cliente. Se antes só se sabia dos dados da compra de uma determinada pessoa, agora se pode saber o que essa pessoa pensa e fala. Isso tem um valor incomensurável quando pensamos em quão rápido os mercados mudam e quão grande é o esforço das empresas para manter o seu cliente. Ficou muito mais fácil traçar todo o perfil de um indivíduo ao tornar tangíveis informações que até então eram inalcançáveis.

Aqui também podemos dar um exemplo do armazenamento desses dados. Ao procurar por dados estruturados de determinado cliente no banco, o analista pode ou receber uma indicação do lugar onde se encontram os e-mails de comunicação deste com a empresa, ou trazer o e-mail junto com os outros dados, o que de fato traria facilidades na análise, mas necessitaria de um poder de processamento muito maior.

Empresas de vanguarda e executivos mais atentos às necessidades de suas companhias já observaram o que Inmon argumenta e isso vem mudando rapidamente. As empresas cada vez mais se dão conta de que não podem desconsiderar esses dados e a informação estratégica contida neles.

Como afirmou Erik Moller, responsável de Information Management da HP Software à CXO, "Os dados não estruturados são quase um tabu para as empresas, que terão sérias dificuldades em compreender totalmente a sua informação de negócio se continuarem a ignorar esta questão". Ainda de acordo com Moller: "os CIO prevêem uma redução significativa na informação não estruturada das empresas ao longo dos próximos três anos. Mas verificamos que a maioria das empresas não tem a noção da quantidade de dados não organizados que existe no seio das suas organizações".

Ele sugere que as companhias atentem para isso e tentem o mais rápido possível automatizar a gestão desses documentos, já que, segundo uma pesquisa da própria HP, 70% dos dados ainda são perdidos por serem não estruturados e não receberem a devida atenção.

Uma explicação para isso é que a princípio pode parecer algo complicado e até mesmo desnecessário, mas vale lembrar que, conforme a análise de Martin

Atherton, analista da consultora Freeform Dynamics, "Podemos debater incessantemente a questão da gestão de informação, mas o que é importante é que as empresas percebam que os recursos de informação ao seu dispor podem ajudá-las a tomar as suas decisões de uma forma mais rápida e eficaz".

Assim, é uma forma valiosíssima a mais de informação, e, da mesma maneira que no início um DW era algo considerado muito dispendioso e talvez dependente de um esforço exagerado, hoje é uma unanimidade entre empresas preocupadas com a estratégia empresarial. O DW 2.0 tende a se mostrar com o tempo tão bem-sucedido quanto o seu predecessor.

2.4.2. Dificuldades

Graças a sua imprevisibilidade já comentada, alguns cuidados devem ser tomados quando lidamos com estes dados. Inmon cita os seguintes problemas: conversas sem importância, terminologia e texto específico ou geral demais.

Como exemplos de conversa sem importância, podemos citar e-mails pessoais trocados entre funcionários. Para uma empresa, de nada importa saber de assuntos entre um funcionário e sua esposa. Grande parte dos e-mails trata disso. Atrapalham muito e devem ser eliminados do nosso procedimento.

O problema da terminologia também traz dificuldades. Pessoas de todo o tipo influenciam os dados não estruturados. Assim, é normal que conforme a idade, classe social, cultural, entre outras diferenças, as pessoas escrevam de maneiras diversas. Uma mesma palavra pode significar coisas diferentes nesse contexto.

Outro problema é que, como o autor do dado provavelmente não tem conhecimento de que seu texto será usado em um DW, não existe a preocupação em utilizar determinado conjunto de palavras. Sendo necessário então, que um analista prepare os documentos de uma forma que seja interpretado pelo computador como deve ser, evitando dados equivocados. Esse processo não é fácil e é uma importante parte do DW.

Deve-se então, buscar a normalização do texto, a ser feito de duas formas: específica e genérica.

Um exemplo desse conceito, conforme Inmon: “O *Tarsus* foi submetido à pressão e *desarticulado*”. Assim seria a versão específica de um texto. Mas se alguém procurar por “osso quebrado”, não trará esse registro como resultado. Procurando por osso não trará *Tarsus* e quebrado não encontrará *desarticulado*. Mas se o texto se encontrar nas duas formas, específico e genérico, o termo *Tarsus* será reconhecido como osso e *desarticulado* será reconhecido como quebrado.

Assim, a partir da frase em questão, dois conjuntos de dados seriam criados: tarsus/osso e desarticulado/quebrado. Este é o segundo grande passo para o tratamento de dados não estruturados, e, sem ele a análise pode perder o sentido, fazendo com que o DW 2.0 se torne um fracasso, não trazendo toda a informação que poderia.

2.5. Visão geral dos dados não estruturados no DW 2.0

Para a carga dos dados não estruturados no banco de dados a fim de torná-los úteis para o DW, são necessários alguns processos. Primeiramente, um processo de ETL, que, como previsível, é diferente do processo de extração e tratamento dos dados estruturados. Este processo será detalhado futuramente, e algumas de suas atividades são:

- ❖ Padronização
- ❖ Remoção de palavras sem utilidade
- ❖ Tratamento de ortografia alternativa
- ❖ Suporte às buscas
- ❖ Uso de Ontologias externas ou internas

Após a realização deste processo, o conteúdo da fonte não estruturada está pronto para ser inserido no banco. Outros tipos de dado fazem parte da estrutura proposta por Inmon. São eles:

- ❖ Taxonomias internas e externas: Uma taxonomia é uma lista de palavras nas quais existe alguma relação entre as palavras. O ambiente textual não estruturado inclui taxonomias que foram criadas internamente (algumas vezes chamadas de “temas”) e taxonomias externas que podem vir praticamente de qualquer lugar.
- ❖ Texto editado e capturado: Texto editado e capturado é o texto que passou através do processo ETL não estruturado e foi colocado em uma base de dados relacional padrão.
- ❖ Links: Links são os dados que amarram os dados não estruturados aos dados estruturados.
- ❖ Apontadores simples: Ocasionalmente os dados não estruturados irão permanecer em outro ambiente e apenas índices de referência para isso serão trazidos para o setor interativo de dados não estruturados do DW.

Estes são detalhados na figura 5:

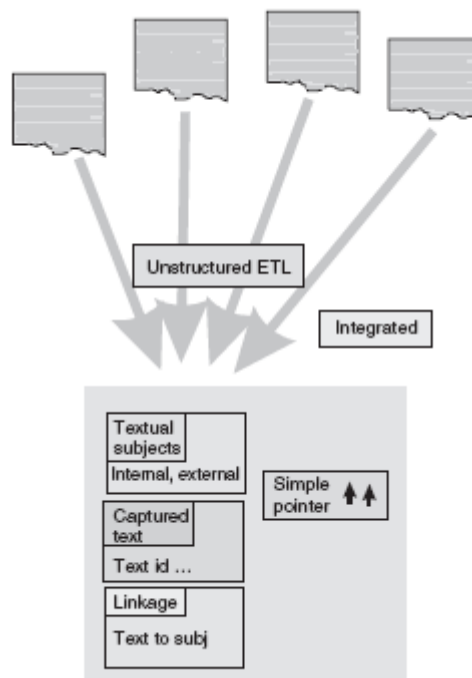


Figura 5 - Inclusão dos dados não estruturados

Outra característica deste tipo de dado é que, como mostrado na figura 6, geralmente apenas duas atividades são ligadas a eles: a carga na base e o acesso.

Não é comum e nem recomendada a atualização dos dados textuais não estruturados. De acordo com Inmon, depois que uma descrição textual do trabalho é realizada, se mudanças necessitam ser feitas ela é completamente reescrita. Dados textuais incrementais ou parcialmente atualizados simplesmente não são um reflexo da realidade.

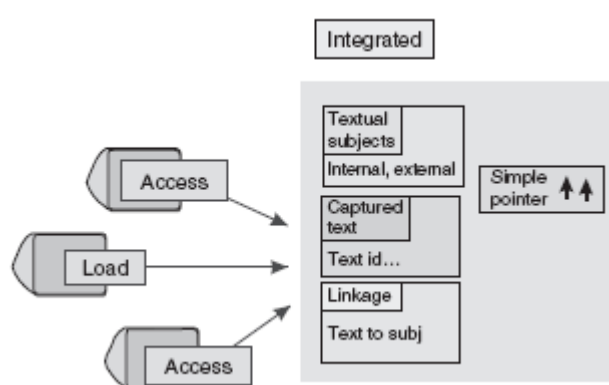


Figura 6 - Acesso aos dados não estruturados na base

2.5.1. Metadados

Metadados são os registros que descrevem os dados. É outra evolução importantíssima do DW 2.0 em relação ao primeiro conceito de Data Warehouse, aonde a princípio, não havia esta preocupação.

Inmon cita diversos casos aonde os metadados são importantes. Como, por exemplo, sua utilidade para os desenvolvedores, que precisam alinhar seu trabalho com o já feito, ou para os técnicos de manutenção que precisa lidar com os problemas do dia-a-dia para manter o Data Warehouse em ordem, e ainda para os usuários finais, que precisam descobrir novas possibilidades de análise.

Bill Inmon propõe ainda um exemplo para mostrar a importância dos metadados. Vê-los como um catálogo de livros em uma grande biblioteca pública.

Como as informações são encontradas na biblioteca pública? As pessoas podem olhar de um lado ao outro de uma prateleira procurando por um livro? Naturalmente sim. Mas é um desperdício de tempo colossal. Existe uma forma muito mais racional para localizar o que você está procurando em uma biblioteca: ir diretamente ao catálogo.

Comparado à busca manual através de todos os livros na biblioteca esta opção é consideravelmente mais rápida e confiável.

Uma vez que o livro desejado é localizado no catálogo, o leitor pode caminhar diretamente para onde o livro está guardado. Ao fazer isso, muito tempo é economizado na localização das informações. Metadados no DW 2.0 executam essencialmente o mesmo papel que o catálogo na biblioteca. Eles permitem ao analista olhar em toda a base da empresa e ver quais análises já foram feitas.

Dados não estruturados no DW 2.0 tem seus próprios metadados. Os metadados para o ambiente dos dados não estruturados são muito diferentes dos usados para o ambiente estruturado. Ontologias são exemplos de ferramentas que podem ser usados desta maneira para o ambiente não estruturado. Outros tipos são:

- ❖ Palavras sem utilidade: São palavras usadas na fala mas que não são importantes no significado do texto. Como: é, são, uns, um, a, o, as, os, quais, cujo, etc.
- ❖ Sinônimos;
- ❖ Homógrafos: Palavras que têm sentido diferente dependendo do contexto. Por exemplo, a diferença do significado da palavra processo para um advogado e para um analista de sistemas.

2.5.2. Envolvimento do Usuário

Se existem dados próximos ao usuário, estes dados são os textuais não estruturados. Dados textuais não estruturados estão presentes no dia-a-dia do

usuário final. Assim, o usuário final é altamente envolvido no processo de inclusão de textos não estruturados no DW 2.0 (INMON; STRAUSS; NEUSHLOSS, 2007).

O usuário final está presente durante quase todo o processo. Isto ocorre pela necessidade que o tratamento de dados não estruturados tem de tratar os dados com conhecimento de negócio. Como exemplo, temos sua participação na especificação de palavras sem utilidade e terminologia.

Geralmente o usuário do negócio tem somente um envolvimento passivo na modelagem dos aspectos estruturados no DW 2.0. Mas o contrário é verdadeiro para os aspectos não estruturados. Por exemplo, o usuário é profundamente envolvido nas especificações da ETL dos dados não estruturados (INMON; STRAUSS; NEUSHLOSS, 2007).

2.5.3. Tratamento

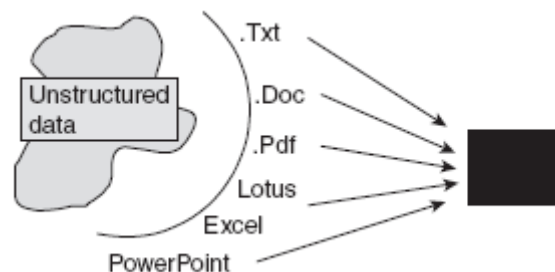


Figura 7 - Tratamento dos dados não estruturados

O primeiro passo no processo de preparação dos dados não estruturados para o processo de análise é a “leitura” do texto pelo sistema. O texto vem em uma considerável variedade de formatos. Os formatos podem precisar ser lidos como entrada. Depois que a fonte original foi lida, o próximo passo é preparar os dados para entrar na base. Esta preparação textual envolve um processo para que os dados possam ser inseridos na base.

Existem várias boas razões do motivo deste processamento, entre elas:

- ❖ Dados não estruturados precisam ser ajustados em um formato relacional;
- ❖ Dados não estruturados precisam ser integrados para que um processo de análise bem sucedido possa ser feito. Se um texto sem análise é simplesmente “jogado” na base, essa análise pode perder o sentido.

2.5.4. ETL

Uma importante decisão deve ser feita no momento anterior à integração dos dados: em que ambiente fazer o tratamento? Pode ser feito no estruturado ou não estruturado. Para que seja feito no lado estruturado é necessária a adaptação dos dados. Pode parecer muito trabalhoso, mas esta adaptação possibilita o uso de tecnologias de análise padrão.

Além disso, como o mundo da análise de dados sempre foi moldado em torno dos dados estruturados, muito já foi gasto em treinamento de usuários e equipes técnicas no ramo de inteligência de negócios visando estes dados. Banco de dados, ETL e processo estatístico já foram desenvolvidos com este propósito. Não faz sentido então, desprezar toda essa tecnologia desenvolvida buscando um novo desenvolvimento caro e trabalhoso. Diante de tudo isso, a escolha se torna simples: o ambiente estruturado é o ideal.

O processo de “integração” textual antes da alocação na base de dados passa por diferentes etapas. Segundo Inmon, uma série de passos é necessária na preparação do texto para incorporação na base de dados e posterior análise no DW 2.0. São elas:

- ❖ Padronização

- ❖ Remoção de palavras sem utilidade
- ❖ Substituição ou concatenação de sinônimos
- ❖ Detalhamento de expressões
- ❖ Criação de temas
- ❖ Uso de Glossários ou Taxonomias externas
- ❖ Radical
- ❖ Ortografia alternativa
- ❖ Internacionalização
- ❖ Suporte a busca direta e indireta

A seguir, teremos um maior detalhamento destes passos.

Padronização

A primeira coisa a se fazer em um texto não estruturado para o processo de tratamento é a sua simples edição. Nisso consiste a uniformização de fonte, forma (letra maiúscula e minúscula) e pontuação. A razão deste passo é fazer com que as buscas não sejam perdidas ou imprecisas por uma letra maiúscula ou outras discrepâncias tipográficas.

Por exemplo, uma busca feita por “DW” deve encontrar o termo “dw”. Caso isso não seja possível, pode ocasionar uma grande perda na análise.

Abaixo um exemplo da eliminação da diferença na forma, fonte e pontuação do texto não estruturado na preparação do mesmo para o processo de análise.

Antes: A ferramenta mais popular para exploração de um *DW* é a *OLAP*.

Depois: a ferramenta mais popular para exploração de um dw é a olap.

Remoção de stop words

O próximo passo é eliminar as palavras que tenham importância para o fluxo da linguagem, mas não sejam significativas para o sentido do texto, sendo assim desnecessárias e até incômodas na análise. São algumas delas:

- ❖ um/uns;
- ❖ e;
- ❖ o/os;
- ❖ a/as;
- ❖ é;
- ❖ para.

Antes: A ferramenta mais popular para exploração de um *DW* é a OLAP.

Depois: ferramenta mais popular exploração *DW* online OLAP.

Substituição de sinônimos

Outro passo que pode ser encontrado na ETL de dados não estruturados para integração na base estruturada é a substituição de sinônimos. Substituição de sinônimos é usada para racionalizar o texto de diferentes terminologias em uma única.

Substituição de sinônimos envolve a substituição de uma única palavra por várias de mesmo significado. Um uso consistente de uma única terminologia, pode ser um importante passo no sentido de garantir a confiança em buscas nos dados não estruturados depois que eles são incorporados ao data warehouse (INMON; STRAUSS; NEUSHLOSS, 2007). Abaixo um exemplo deste processo.

Antes: A ferramenta mais popular para exploração de um *DW* é a OLAP.

Depois: O tipo de software mais usado para análise de um *DW* é o OLAP.

Concatenação de sinônimos

Uma alternativa à substituição de sinônimos é a concatenação de sinônimos. Neste processo, ao invés de substituir sinônimos por uma palavra padrão, a palavra padrão é inserida próxima ao, ou concatenada com, todas as ocorrências de outras palavras que tenham o mesmo sentido, seus sinônimos. Ou seja, todas as possibilidades que o usuário tem de se referir àquela palavra, são unidas, fazendo assim com que a busca encontre qualquer expressão referente a um assunto em questão.

Antes: A ferramenta mais popular para exploração de um *DW* é o OLAP.

Depois: O tipo de software A ferramenta mais popular usado para exploração análise de um *DW* é o OLAP.

Detalhamento

O detalhamento é o contrário da concatenação e substituição de sinônimos. O detalhamento é usado para definir o significado de palavras ou frases que possam significar mais de uma coisa dependendo do contexto em que são inseridas. O verdadeiro significado substitui ou sobrepõe a palavra ou frase que aparece no texto.

Antes: A ferramenta mais popular para exploração de um *DW* é o OLAP.

Depois: A ferramenta mais popular para exploração de um *Data Warehouse* é o Processo Analítico em Tempo Real.

Criação de Temas

Uma das possibilidades, depois de ser feita a integração do texto, é produzir um “cluster” do texto. Clusterização textual é uma proposta de produção de “temas”. Na clusterização textual, palavras e frases são agrupadas logicamente baseadas no número de ocorrências das palavras e da sua proximidade. Pode levar também a uma ontologia ou um glossário, sendo chamado glossário ou ontologia interna por ter sido criado a partir do texto.

Ontologias e ontologias externas

Da mesma maneira que ontologias ou glossários internos são úteis, ontologias e glossários externos também podem ser. Ontologias e glossários externos podem representar qualquer coisa. Eles podem ser usados para impor uma estrutura ao texto. O texto pode ser lido do sistema e uma comparação pode ser feita para determinar se o texto pertence ou é relacionado de outra maneira à uma ontologia ou glossário externo, fazendo assim com que seja decidido se o texto, ou determinadas frases são úteis ou não ao DW.

Redução ao radical

A redução ao radical é outro passo na integração do texto e preparação para análise textual. Radical é a raiz grega ou latina da palavra. A redução ao radical é importante se o objetivo é encontrar palavras com origem em comum. Por exemplo, caso se queira achar qualquer palavra relacionada ao termo “corrida” e se procure literalmente. Perderemos palavras como corredor e correr. Caso seja usado o radical para a busca, isso não acontece.

Ortografia alternativa

Caso as buscas sejam efetivamente feitas, deve-se pensar na necessidade de escrita alternativa de algumas palavras, ou mesmo ortografias erradas. Em alguns documentos rápidos como e-mail, ocorrem erros de ortografia. Caso não se queira perder isso, é interessante o cuidado com esse passo.

Internacionalização

A possibilidade de um processo de internacionalização deve ser considerada em alguns tipos de documento, dependendo do negócio. Para multinacionais, por exemplo, que podem lidar com documentos em diversas línguas, é algo extremamente útil.

Busca direta

Outra importante possibilidade da análise textual é a habilidade de dar suporte a diferentes tipos de busca. A integração de texto precisa definir um estágio para isto. Um tipo de busca que precisa ser suportado é a busca direta. Busca direta é aquela feita por ferramentas de sites de busca como Google ou Yahoo. Um argumento é passado ao mecanismo de busca que procura por qualquer ocorrência disso.

Busca Indireta

Outro tipo de busca é a busca indireta. Aonde parâmetros de busca são igualmente passados ao mecanismo de busca, mas em uma busca indireta, a procura não é uma feita no próprio argumento.

Nesta procura, podem ser utilizadas ontologias, que relacionarão a busca à assuntos que tem algum tipo de associação à palavra buscada.

Terminologia

O processo de terminologia é uma das etapas mais difíceis para a análise. Este problema acontece porque nossa linguagem é em terminologia. Cada grupo de pessoas se refere a um assunto de uma maneira, dependendo de classe social, cultural ou profissional.

Inmome dá o exemplo do corpo humano. Para qualquer parte do corpo humano pode haver em torno de 20 diferentes maneiras de se referir àquela parte. Um doutor usa uma terminologia. Outro doutor usa outra terminologia. Uma enfermeira usa outra. Todos estes profissionais estão falando da mesma coisa. Entretanto, estas pessoas estão falando linguagens diferentes.

Se o processo analítico é feito no texto, precisa haver uma resolução da terminologia. A base de dados de palavras e frases deve ser armazenada conjuntamente de modo específico e genérico.

Por fim, o dado textual deve conter a palavra específica original usada pelo doutor ou enfermeira para definir a parte do corpo humano, e a base de dados textual deve conter um termo que será entendido por toda a comunidade de analistas.

O processo de terminologia é extremamente necessário para que seja possível fazer um processo de análise textual específico.

Dados semi estruturados/valor

Dados não estruturados aparecem em todos os tipos. A forma mais simples de dados não estruturados é de texto em um documento. Aonde existe texto em um documento, as palavras e frases não têm ordem ou estrutura. Um documento não estruturado é apenas isso, um documento não estruturado. Mas existem outras formas de documentos textuais. Em alguns casos, o autor do documento pode ter dado uma estrutura ao documento. Um exemplo simples de um documento com uma estrutura interna é um livro de receitas. Dentro de um livro de receitas existem

muitas receitas. É um documento. Mas dentro do documento existem inícios e fim explícitos. Uma receita acaba e outra começa.

Habitualmente também é necessário mapear as estruturas implícitas do livro sobre a base de dados textual analítica. Em alguns casos, isso é muito fácil e óbvio de se fazer. Em outros casos não é tão evidente. Outra forma de dado não estruturado usado de exemplo por Inmon é o de um currículo. Apesar de não ter uma estrutura convencional para banco de dados, existe uma espécie de molde nestes documentos. Por exemplo, em todos os currículos se encontram entradas comuns, como nome, endereço e educação.

Tendo conhecimento deste molde, pode-se adaptar o texto de modo a conseguir algo como “nome – João da Silva”. Isto propicia ao sistema o reconhecimento de que o campo nome é importante e a pessoa é João da Silva. Com isto, o texto pode ser lido e as palavras apanhadas simbolicamente, não literalmente. Esta capacidade de sentido simbólico é de grande valia para o desenvolvimento de análises textuais.

Colocação na base estruturada

Quando o texto não estruturado está pronto para o processamento analítico, o texto é alocado em uma base de dados relacional. A base de dados relacional pode ser acessada e analisada por diferentes ferramentas, como ferramentas de BI.

Link de dados estruturados/não estruturados

Depois que a base de dados não estruturada é criada, é feita então uma ligação com a base de dados estruturada, para formação do DW 2.0 na empresa.

3. PLANEJAMENTO

Desnecessário falar sobre a importância do planejamento antes da execução de qualquer sistema de informação. Principalmente quando o assunto é novo e sem muitos exemplos, como a extração de informações a partir de documentos não estruturados para um Data Warehouse.

O planejamento, segundo Reynolds capacita administradores a direcionar os esforços e recursos da organização para o alcance de objetivos específicos.

É então um passo que evita o desperdício de tempo e recursos no desenvolvimento de um sistema que depois se mostrará insuficiente para o atendimento dos requisitos do negócio.

3.1. Proposta do Trabalho

Conforme ilustrado na figura 8, a visão geral do trabalho será:

1. Estudo de quais partes do acórdão nos serão úteis.
2. Desenvolvimento do plugin para extração de dados a partir desta análise.
3. Preparação dos campos estruturados a serem integrados.
4. Extração dos dados não estruturados e integração com os dados vindos das tabelas estruturadas, utilizando o Kettle.
5. Desenvolvimento dos Data Marts.
6. Informações obtidas a partir da extração da informação dos acórdãos.

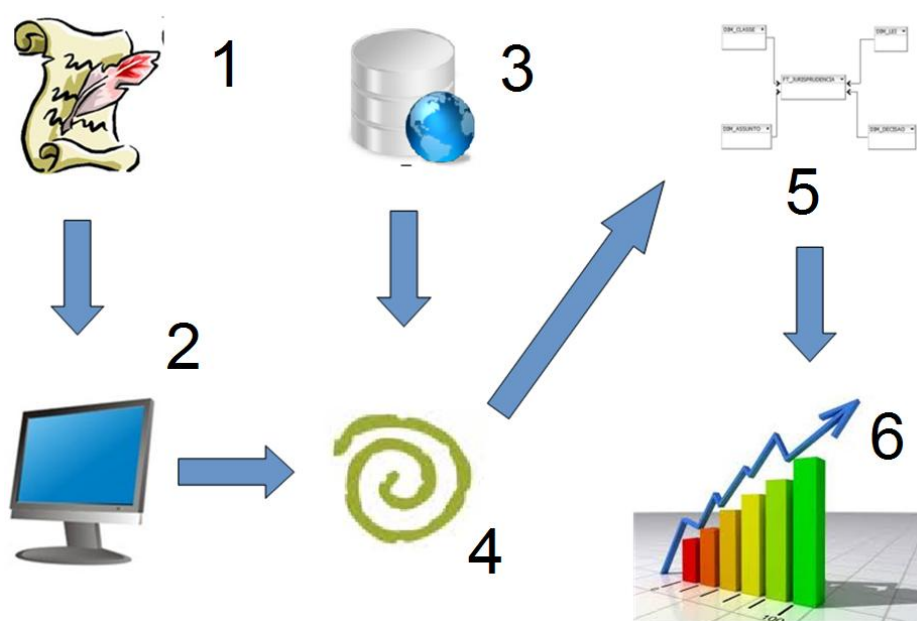


Figura 8 - Estrutura geral do projeto

Segue um maior detalhamento de cada uma destas etapas.

3.1.1. Análise do Acórdão

A análise do negócio é sempre uma das partes principais, senão a mais importante, de um Data Warehouse. O sucesso depende, em grande parte, da correta identificação dos problemas e de como resolvê-los através das ferramentas que temos em mãos.

Nosso problema é a falta de dados para apoio da jurisprudência, e nossa fonte para a resolução desta carência é a extração de informações a partir do acórdão.

Conforme já mencionado, um acórdão tem a seguinte estrutura: dados do processo, ementa, relatório, voto e decisão.

A parte de relatório do processo é, por vezes, bastante extensa, o que dificulta a consulta e análise. Justamente nesta parte se encontra um dado chave para o nosso projeto, a lei usada na decisão.

Tendo como ideia deste trabalho possibilitar o acesso às estatísticas do emprego da jurisprudência ou, até mesmo, facilitar o acesso direto à conclusão de

determinado processo, o tratamento adequado para este documento é a obtenção de um resumo da decisão, trazendo o voto do magistrado, e em qual lei foi baseada esta sentença.

Hoje em dia já são armazenados no sistema a Classe e o Assunto do processo, responsáveis por definir o tipo do processo. Ao unirmos a isso a lei utilizada e a decisão tomada a partir dela, possibilitaremos a criação de estatísticas sobre a Jurisprudência.

3.1.2. ETL do documento não estruturado

Depois de decidido, na análise do acórdão, quais os pontos necessários para que o documento se ajuste às necessidades do trabalho, foram definidos os passos da extração de documentos não estruturados proposta por Inmon que melhor se adaptam ao caso do tratamento do acórdão.

A partir desta decisão, foi desenvolvida uma extensão utilizando a linguagem de programação Java. Este futuramente será integrado ao Kettle, software para realização de Integração de dados já citado.

Com a definição de quais passos propostos na teoria do DW 2.0 poderiam ser usados neste trabalho, identificamos quais características deveriam ser desenvolvidas no plugin.

- Primeiramente a padronização, fazendo com que as palavras fiquem todas com a mesma fonte e com letras minúsculas, facilitando as buscas no texto.
- O segundo passo é a remoção de palavras e termos sem utilidade. Como a ideia é deixar o documento útil para ser usado no apoio à jurisprudência, na análise foi decidido por buscar a decisão do magistrado e a lei na qual essa decisão foi baseada.
- A seguir, a substituição de sinônimos, passo utilizado nas decisões presentes nos documentos, que podem usar diferentes palavras para um mesmo significado.

Buscando a integração com o DW atual, será necessário também o número do processo, de forma que, a partir deste código, conseguiremos ligar o acórdão ao seu processo já cadastrado na base de dados.

Sendo assim, durante o tratamento feito no acórdão, extrairemos o número do processo, a decisão tomada e a lei. A decisão e a lei se tornarão duas dimensões com chaves em colunas da nossa tabela e o número do processo foi usado para ligarmos estas informações ao seu respectivo processo.

3.1.3. Integração ao banco de dados estruturado

Para o projeto de integração, é necessário o entendimento do padrão atual do DW.

Podemos, hoje, gerar diversas estatísticas baseadas em fases ou características do processo. Para isso, existem vários Data Marts, como: Andamento Médio dos Processos, Movimentação do Processo e Tempo Médio do Processo.

No sistema atual, cada Data Mart possui somente as chaves de dimensões consideradas importantes para identificação de processos semelhantes e uma coluna quantitativa. Isso possibilita a criação de estatísticas.

Como exemplo, em uma tabela fato responsável por informar estatísticas sobre o tempo do processo, utilizaríamos chaves relativas ao tema do processo (Dim_Classe e Dim_Assunto), ao local aonde o processo se encontra (Dim_Tribunal) e ao tempo (Dim_Tempo). A figura 9 mostra um exemplo de Data Mart do Data Warehouse atual, com todas as dimensões tendo como fonte, os dados estruturados.

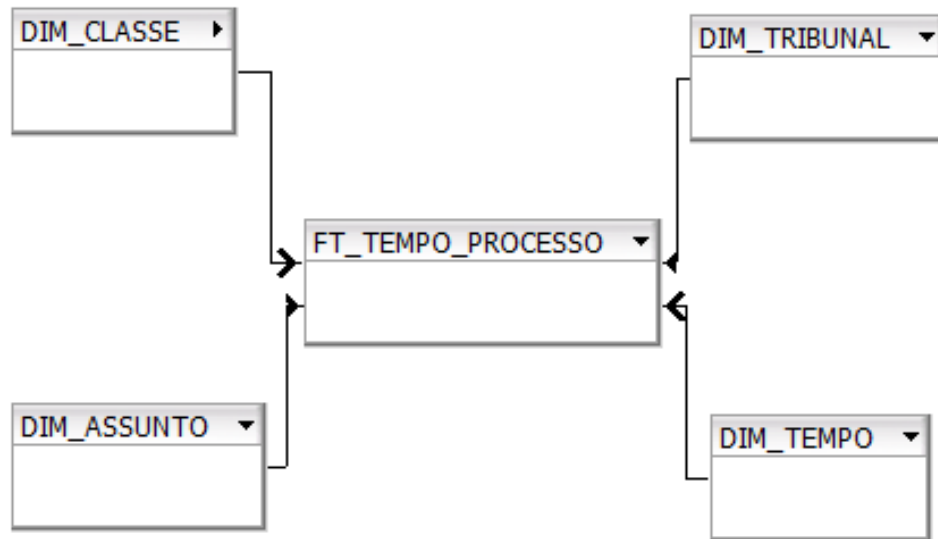


Figura 9 - Exemplo de estrutura no Data Warehouse atual

Após análises efetuadas no documento, vimos que serão necessárias duas novas dimensões para o nosso projeto, visto que pretendemos criar estatísticas baseadas na Lei e na Decisão judicial.

Baseado nesse esquema, planejamos nossa tabela para os dados não estruturados.

A tabela fato Jurisprudencia terá três chaves comuns às demais tabelas semelhantes do banco, essas tem como fonte os dados estruturados, são elas: a Dimensão Classe, a Dimensão Assunto e a Dimensão Tempo. Junto a isso, a tabela terá também ligação com as duas novas dimensões, provenientes dos dados não estruturados, a Dimensão Lei e a Dimensão Decisão. Será possível então, através desta, verificar quantos processos de mesma classe e assunto foram decididos baseados na mesma lei e tiveram a mesma decisão. Detalharemos melhor as dimensões a seguir.

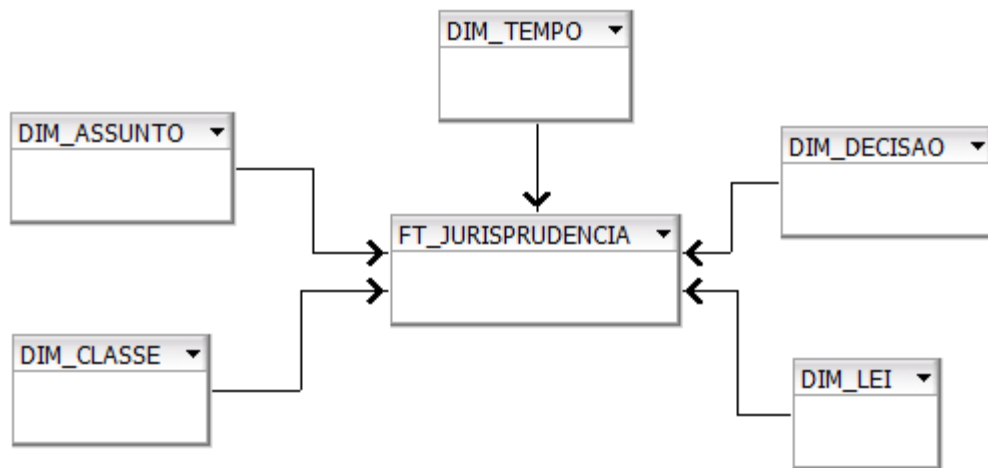


Figura 10 - Tabela fato jurisprudencia, criada no projeto

- Dimensão Classe

Dimensão já presente no banco de dados atual. Será integrada aos dados do acórdão pelo código do processo. As classes são determinadas pelo Conselho Nacional de Justiça e, junto com o assunto, informam a natureza do processo.

- Dimensão Assunto

Assim como a Dimensão Classe, esta já existe no DW atual e sua origem são os dados estruturados. Indica do que se trata o processo. Responsável no sistema por apontar processos similares para a geração de estatísticas.

- Dimensão Lei

Informação impossível de ser obtida atualmente a partir dos dados estruturados será preenchida a partir da extração nos acórdãos.

Indicará quais leis foram utilizadas para a tomada de decisão. Junto com a decisão é a principal inovação deste trabalho e o que tornará possível o seu objetivo.

- Dimensão Decisão

Informará a decisão do órgão colegiado, conforme descrito no acórdão. Também preenchida a partir dos dados não estruturados. Indicará, junto com a dimensão lei, a resolução do processo e os motivos.

3.1.4. Análises

Depois de tratados e indexados, os documentos estão prontos para as funcionalidades do DW.

Nesta etapa serão geradas as estatísticas relativas ao uso da jurisprudência e poderemos também efetuar a busca em uma decisão de um processo específico.

4. INTEGRAÇÃO DOS DADOS

Depois do planejamento, é o momento de testar e avaliar o sistema. Neste capítulo apresentaremos os resultados gerais, quanto às fontes de informação utilizadas, cargas de dados nas dimensões e outras tabelas e por fim analisaremos as informações obtidas.

4.1. Fontes de informação

Neste projeto duas fontes de informação foram utilizadas, os dados estruturados já utilizados no atual Data Warehouse e os dados não estruturados provenientes dos Acórdãos.

4.1.1. Acórdãos

Os acórdãos utilizados foram extraídos do DVD Jurisprudência Catarinense 2009, obtido no Tribunal de Justiça de Santa Catarina. Como não existe, no DVD, a possibilidade de extração dos arquivos, os mesmos foram copiados e salvos em arquivos de textos para que pudessem servir ao projeto. Foram utilizados diversos acórdãos de diferentes tipos, de modo que se pudesse analisar individualmente a correta integração e geração de estatísticas.

Conforme citado no planejamento, as informações extraídas dos acórdãos foram o código do processo, a lei utilizada e a decisão final.

4.1.2. Dados estruturados

A fonte dos dados estruturados foram dados de teste do banco de dados já existente no Sistema Judiciário de Santa Catarina. Conforme definido no planejamento, os dados estruturados buscados foram a classe e o assunto de cada processo.

4.2. Integração com o DW

As duas fontes de dados citadas no capítulo anterior são o ponto inicial do processo. Após a entrada no processo de ETL, os acórdãos passam pelo tratamento para extração das informações necessárias. Este tratamento é feito através do plugin desenvolvido para o Kettle. Os campos da lei utilizada e da decisão tomada, após a extração feita com a extensão desenvolvida, formarão duas dimensões: a Dim_Lei e a Dim_Decisao.

Depois de obtido o código do processo, através dele será realizada a integração com o banco estruturado. A partir deste código, será feita a junção para a busca da classe e do assunto do processo.

Neste ponto do trabalho já teremos todas as informações necessárias para o preenchimento da tabela fato.

4.3. Extensão Desenvolvida

Neste capítulo detalharemos o Plugin criado para o Kettle, responsável pela busca e extração das informações a serem obtidas nos dados não estruturados.

4.3.1. Tratamento de Acórdão

O plugin desenvolvido para o projeto foi criado visando o tratamento dos dados do Acórdão. Seu processo é simples: após a entrada do documento no fluxo, realizada através do Kettle, é feita uma busca no documento, procurando por palavras chaves previamente definidas de acordo com a análise dos acórdãos. Estas palavras definirão as informações que buscamos. Por exemplo: ao encontrarmos “nego o provimento”, saberemos que a decisão foi de negar a Apelação.

Quando alguma das palavras é encontrada, seu significado no caso da decisão ou a própria palavra no caso da lei é guardada e colocada no fluxo da ETL.

Como seu procedimento é totalmente interno e padrão, não existe uma tela de interface com o usuário. Na figura 11 vemos o exemplo de seu uso.

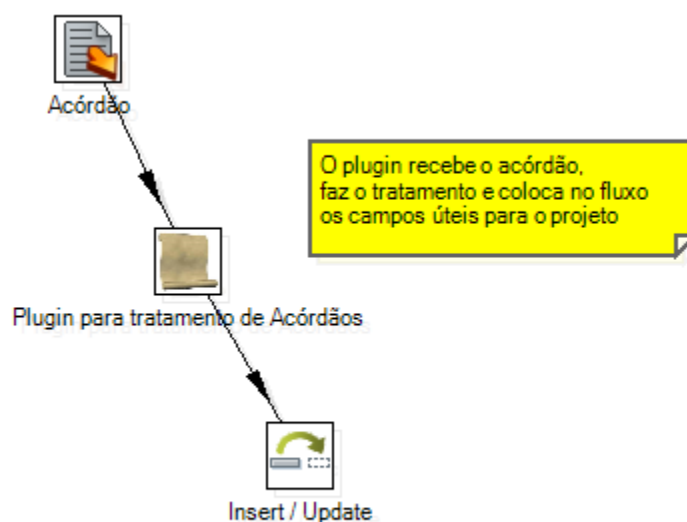


Figura 11 - Utilização do plugin desenvolvido

Para um melhor entendimento, analisaremos o código desenvolvido.

Primeiramente é necessária a explicação de que o software Kettle envia para o plugin cada linha do acórdão individualmente, por este motivo a análise foi então feita linha a linha.

Conforme já vimos no capítulo sobre Acórdãos, o início do documento, aonde nos é informado o número do processo, tem o seguinte padrão:

Apelação nº: 2934856

Por aparecer no início do documento, o número da apelação é a primeira informação a ser buscada. Para isto, primeiramente conferimos se esta palavra existe no documento, através da função `lastIndexOf`, nativa da linguagem Java. Caso o conjunto de caracteres exista, esta função retornará "-1". Após a confirmação da sua existência, buscaremos a informação do número do processo, o que será feito através da função `substring`, retornando os caracteres posteriores à "Apelação nº:", ou seja, o número do processo, que no nosso exemplo seria "2934856", esta informação será então colocada no fluxo já apresentado e não será mais buscada nas próximas linhas recebidas pela nossa extensão.

```
int apelacao = acordao.lastIndexOf("apelação nº");
```

```
if (apelacao != -1){  
    acordao = acordao.substring(str.indexOf(":"));
```

A seguir, buscaremos nas próximas linhas recebidas o número da Lei. Em todos os acórdãos analisados foi observado o mesmo padrão nesta informação, o uso da expressão "Lei n.º" antes do número da lei utilizada. Seguindo o exemplo anteriormente citado, receberíamos a seguinte linha:

da Lei n.º 8.429/92, bem como ao pagamento de honorários

Um procedimento muito semelhante ao utilizado para a extração do número da apelação é utilizado, com a diferença de que como podemos ter outras

informações após o número da lei, a função substring retorna o conjunto de caracteres desde o final de “Lei n.º” até o próximo espaço vazio.

```
int lei = str.lastIndexOf("lei n.º");
if (lei != -1){
    substr = str.substring(str.indexOf("º")+ 2, str.indexOf(" ", str.indexOf("º")+ 2));
```

Depois de já termos preenchido o número do processo e a lei utilizada, procuraremos nas próximas linhas por uma série de palavras que podem determinar a decisão do juiz e as enquadraremos em um dos três tipos: provimento concedido, provimento negado ou provimento parcial.

```
int provimento = str.lastIndexOf("provimento");
int nego = str.lastIndexOf("nego");
int negado = str.lastIndexOf("negado");
int denego = str.lastIndexOf("denego");
int parcial = str.lastIndexOf("provimento parcial");
int parcial2 = str.lastIndexOf("parcial provimento");
int concedido = str.lastIndexOf("concedido");

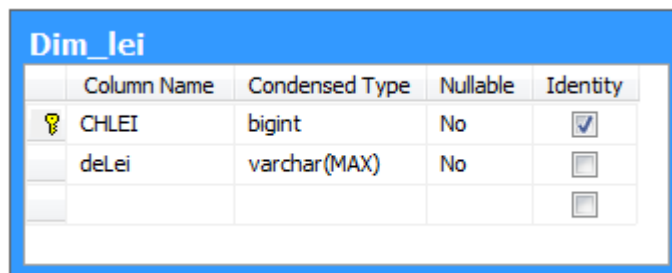
if ((provimento != -1){
    if ((nego != -1) || (denego != -1) || (negado != -1)){
        str = "provimento negado";
    }else if ((parcial != -1) || (parcial2 != -1)){
        str = "provimento parcial";
    }else if (concedido != -1) {
        if (concedido != -1){
            str = "concedido";
```

Depois de recebidas estas informações, elas são colocadas em suas respectivas variáveis e inseridas no fluxo do Kettle visto na figura 11 para inserção nas tabelas.

4.4. Processo de ETL

4.4.1. Dimensão Lei

Primeiramente vamos falar da Dimensão Lei. Denominada Dim_lei no nosso DW. Conforme mostrado na figura 12, é uma dimensão extremamente simples. Seus campos são o CHLEI, a chave da dimensão, e a descrição da lei. Como podemos ver na figura, o campo CHLEI é preenchido diretamente pelo banco de dados.



Column Name	Condensed Type	Nullable	Identity
CHLEI	bigint	No	<input checked="" type="checkbox"/>
deLei	varchar(MAX)	No	<input type="checkbox"/>

Figura 12 - Dimensão Dim_lei

Na figura 13 podemos ver como o campo de descrição da lei, chamado deLei é preenchido.

Primeiramente os dados não estruturados entram no fluxo de ETL, como pode ser observado no passo denominado “Acórdão”. Então o plugin desenvolvido para este trabalho é executado. O plugin vai percorrer o documento procurando as informações já citadas. A seguir é excluída qualquer linha vazia que possa aparecer, etapa executada apenas por segurança. A seguir, o passo “Row Flattener” transforma as linhas recebidas do plugin de tratamento de acórdão em colunas. Uma delas será a coluna “Lei”, com a qual a dimensão será preenchida, como mostrado em destaque.

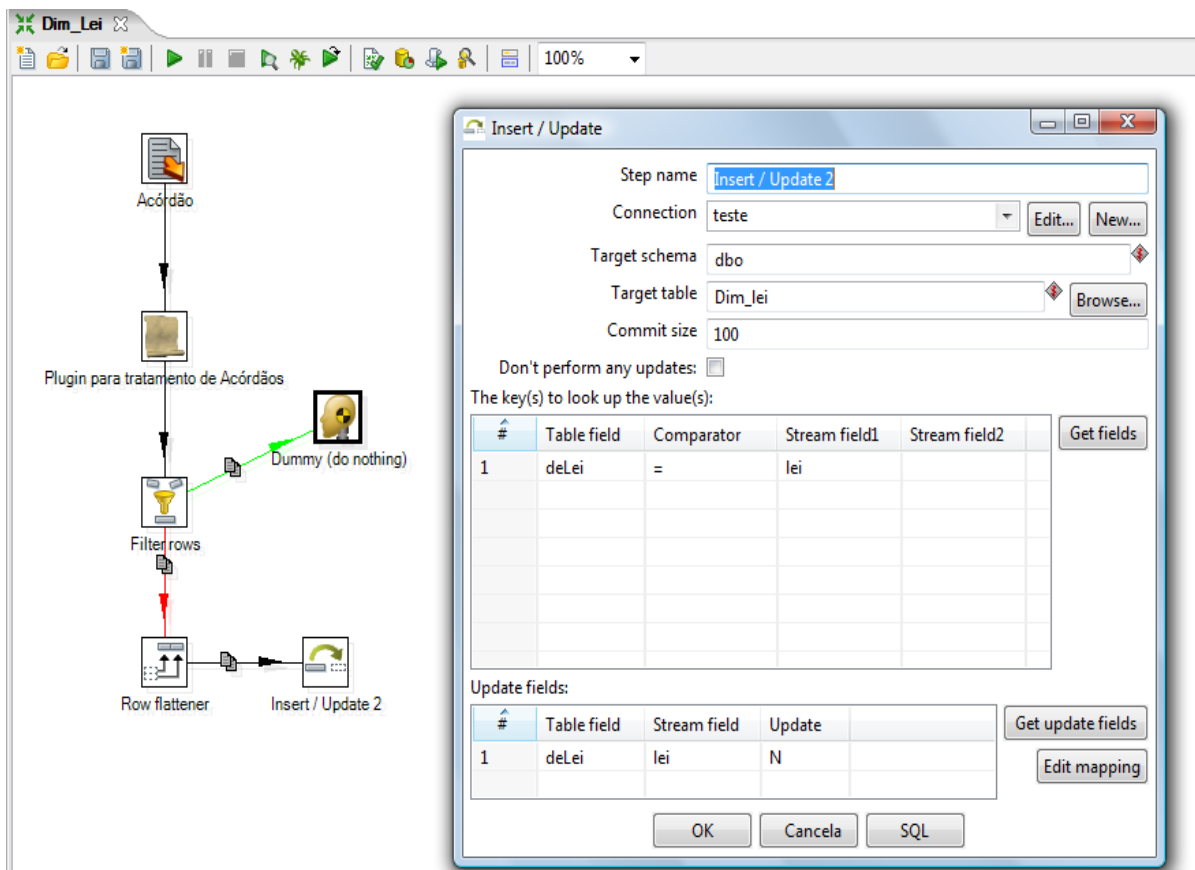


Figura 13 - ETL da Dimensão Dim_Lei

4.4.2. Dimensão Decisão

A Dimensão de Decisão é bastante similar à de Lei, tanto na sua estrutura quanto no seu tratamento.

Dim_decisao				
	Column Name	Condensed Type	Nullable	Identity
	CHDECISAO	bigint	No	<input checked="" type="checkbox"/>
	deDecisao	varchar(MAX)	No	<input type="checkbox"/>
				<input type="checkbox"/>

Figura 14 - Dimensão Dim_decisao

A dimensão é formada por uma chave e pela descrição da decisão. Esta, como a descrição da lei, também será obtida a partir do tratamento do acórdão. Mas, enquanto na descrição relativa à Lei ela simplesmente é inserida na tabela, na descrição da decisão é realizado um tratamento pelo plugin. A dimensão receberá o significado da decisão, e não uma transcrição do Acórdão. Este passo já foi exemplificado na seção sobre o desenvolvimento do Plugin, mas comentaremos novamente usando como apoio a Figura 15 que nos mostra os dados da tabela.

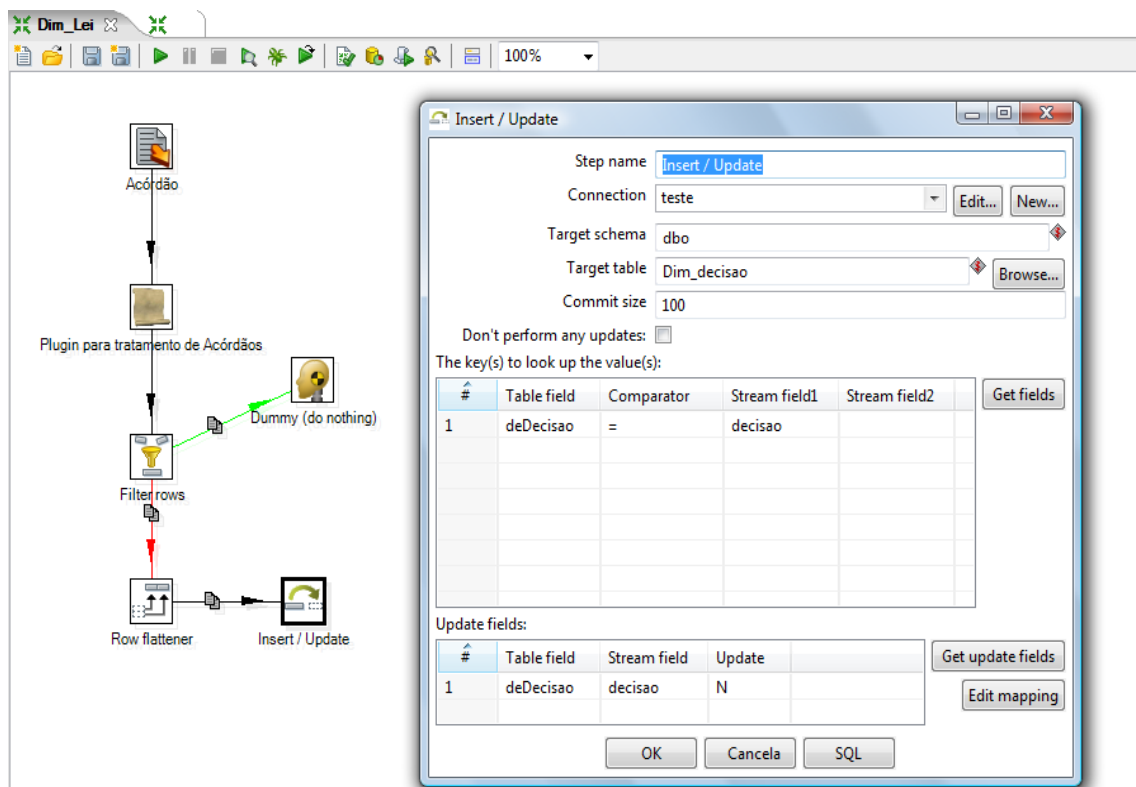


Figura 15 - ETL da Dimensão Dim_decisao

Caso conste no acórdão decisões como: “nego o provimento”, “denego”, “provimento negado”, todas elas passarão para a base como “Provimento Negado”. Assim como “dou provimento” entrará na tabela como “Provimento Dado”.

Esta adaptação tem por motivo a criação de estatísticas, de modo que o importante é termos o resultado da decisão, independente de como foi transcrita.

4.4.3. Jurisprudência

A tabela fato Jurisprudencia será a fonte das estatísticas a serem geradas. Tem como colunas as chaves das dimensões e um campo totalizador.

Jurisprudencia				
	Column Name	Condensed Type	Nullable	Identity
🔑	CHLEI	bigint	No	<input type="checkbox"/>
🔑	CHDECISAO	bigint	No	<input type="checkbox"/>
🔑	CHCLASSE	bigint	No	<input type="checkbox"/>
🔑	CHASSUNTO	bigint	No	<input type="checkbox"/>
	totalProces...	int	No	<input type="checkbox"/>
				<input type="checkbox"/>

Figura 16 - Tabela fato Jurisprudencia

Com os dados desta tabela poderemos ver, por exemplo, quantos processos foram negados do assunto Furto, da classe Habeas Corpus que usaram determinada lei para esta decisão. Para ilustrar melhor o acesso à essa tabela, temos um exemplo de consulta SQL na figura 17.

```
SELECT deAssunto, deDecisao, deLei, totalProcessos FROM Jurisprudencia J
JOIN Dim_lei L ON J.CHLEI = L.CHLEI
JOIN Dim_decisao D ON J.CHDECISAO = D.CHDECISAO
JOIN Dim_assunto A ON J.CHASSUNTO = A.CHASSUNTO
JOIN Dim_classe C ON J.CHCLASSE = C.CHCLASSE
WHERE A.CHASSUNTO = 2042
```

Figura 17 - Consulta SQL na tabela fato

Esta consulta trará como resultado o número de processos de determinado assunto que seguirem a mesma lei e com isso obtiveram a mesma decisão. Com ele podemos comparar quantos processos semelhantes obtiveram decisões contrárias.

5. RESULTADOS

Após a análise dos acórdãos, o desenvolvimento do plugin, o estudo para integração com os dados estruturados e a criação de tabelas, foi executada a carga dos acórdãos a partir de dados reais obtidos no DVD Jurisprudência Catarinense 2009. Foram extraídos deste DVD cerca de 200 acórdãos, a fim de testar a integração com o DW atual e possibilitar o descobrimento de algum padrão nas decisões dos magistrados. Quanto ao preenchimento das dimensões, o processo obteve grande sucesso. A única observação a ser feita é a dificuldade em relação à identificação das palavras que identificam a decisão, visto que, diferentes termos são utilizados dependendo do relator do acórdão. O tratamento destes casos, previsto por Inmon e descrito na seção sobre tratamento dos dados não estruturados no item “Substituição de sinônimos”, foi considerado no desenvolvimento do plugin. Mesmo assim, como são expressões que mudam constantemente, será preciso um melhor tratamento quando a carga de acórdãos for consideravelmente maior. Na figura 18 podemos ver os dados das dimensões.

	CHLEI	deLei
1	1	8.429/92
2	3	8.987/95
3	6	44,
4	7	155
5	8	157
6	9	311

	CHDECISAO	deDecisao
1	1	provimento parcial
2	6	provimento negado
3	9	concedido
4	11	negado

Figura 18 - Exemplos de dados das Dimensões

A integração dos dados também foi bem sucedida. Com o número do código do processo as chaves de Classe e Assunto foram ligadas de maneira correta aos processos dos acórdãos, assim como as dimensões da lei utilizada e da decisão tomada. Na carga dos acórdãos e na criação de estatísticas foi observado um problema. Os acórdãos variam muito de assunto e classe, e como a base destes documentos não tinha uma organização segundo o tipo de processo, foi extremamente complicado o acompanhamento de processos semelhantes.

Mesmo com estas dificuldades, foi possível a análise de processos de Habeas Corpus relativos à prisões pelo assunto tráfico de drogas. Através da análise da estatística desses processos notamos uma clara tendência quanto à jurisprudência deste caso específico, como podemos observar no gráfico da figura 19.

Habeas Corpus para suspeitos de tráfico de drogas

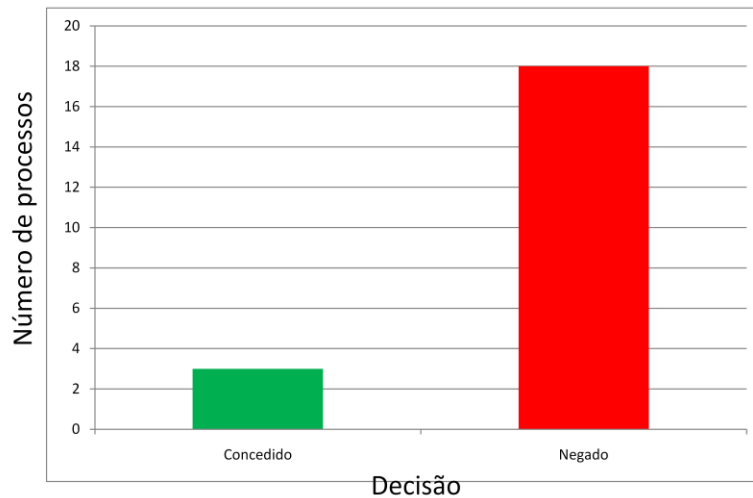


Figura 19 - Gráfico do resultado de pedidos de Habeas Corpus para suspeitos de tráfico de drogas

Podemos notar um claro padrão quanto às decisões. De cada 21 acórdãos do tipo citado solicitando a suspensão da prisão, 18 foram negados. Portanto, este caso é um exemplo de que o objetivo do projeto foi atingido, ou seja, a correta extração e integração de dados não estruturados possibilitando a descoberta de padrões nas decisões da justiça do segundo grau.

Outros padrões também foram descobertos. Em 2009 houve diversos casos em que os professores não receberam determinados benefícios por estarem de licença devido a motivos de saúde. Todos os acórdãos analisados consideraram isso ilegal e decidiram favoravelmente aos professores, como podemos ver na figura 20.

Mandado de Segurança para professores de licença

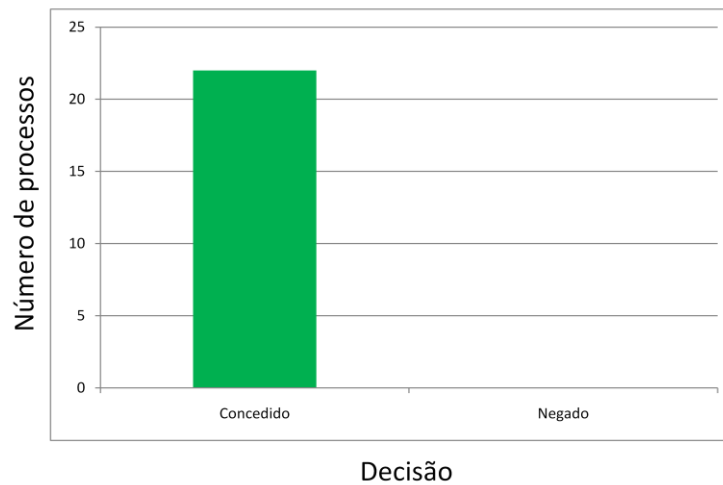


Figura 20 - Gráfico de mandado de segurança para professores de licença

Além destes dois usos da jurisprudência na decisão dos magistrados, podemos também observar na figura 21 o mesmo padrão na decisão quanto à transferência de ações quando da compra de uma empresa de telefonia estatal por uma empresa privada.

Ações de empresa de telefonia compradora de estatal

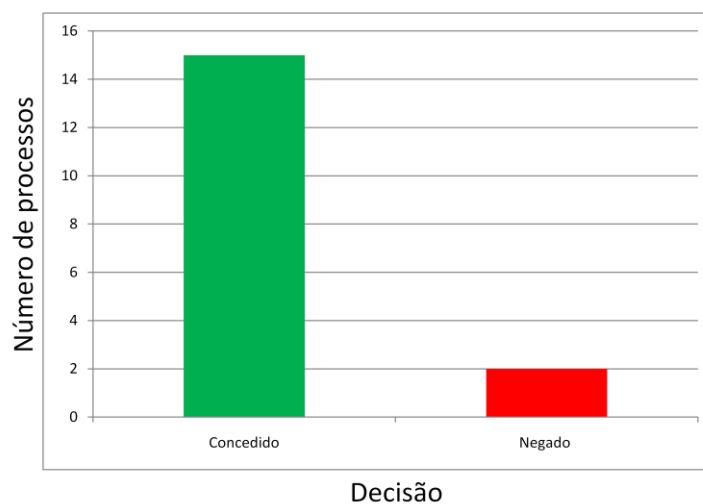


Figura 21 - Gráfico dos processos quanto à mudança na equivalência de ações após venda de companhia

Estes três casos nos apresentam claramente a contribuição que poderia ser dada por este trabalho. Um magistrado ao analisar estatísticas que mostrassem que todos deram ganho de causa a determinado lado, como nos casos mostrados, poderia tomar uma decisão em um tempo muito menor, o que traria uma grande contribuição à rapidez no andamento dos processos.

Outros casos mais específicos também foram encontrados, como a solicitação de indenização por danos morais quando o cliente foi inscrito em serviços de proteção ao crédito de forma indevida mostrado na figura 22 ou o ressarcimento de danos em acidentes de automóveis na figura 23.

Solicitação de danos morais por inscrição no PROCON

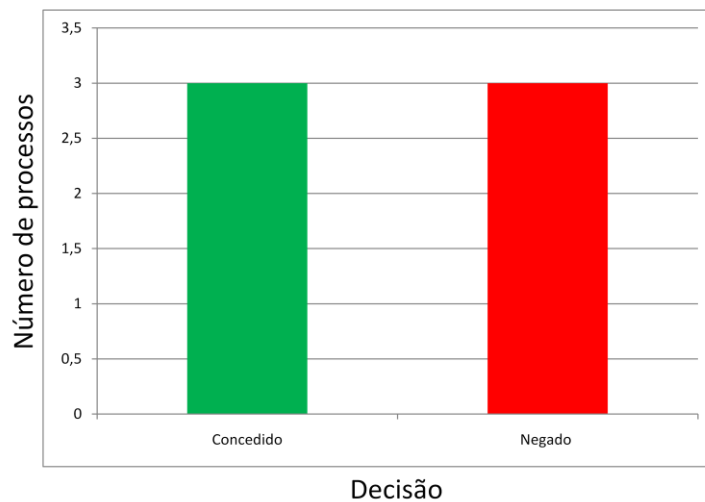


Figura 22 - Gráfico relativo à ações de dano moral por inscrição indevida em serviços de proteção ao crédito.

Ressarcimento de danos pedido por seguradoras de veículos

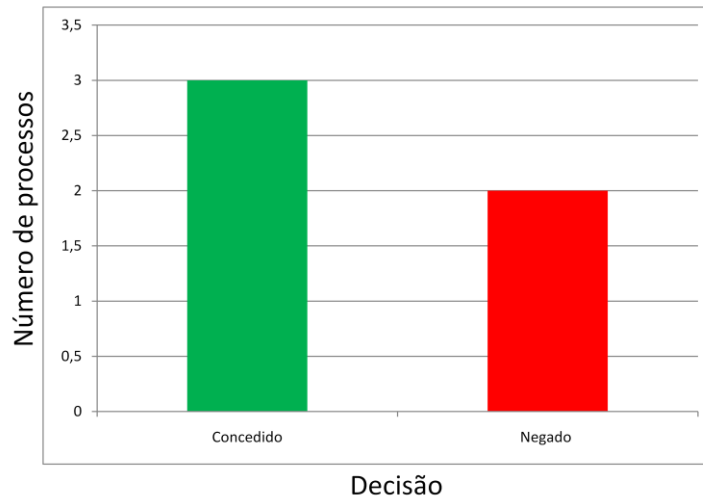


Figura 23 - Gráfico dos processos quanto ao ressarcimento de danos em acidentes automotivos

Como, apesar de semelhantes, estes processos guardam muitas particularidades individuais, fica mais difícil o encontro de um padrão nas decisões; Com as estatísticas podemos observar também estes casos, onde o uso da jurisprudência fica prejudicado e uma análise maior é necessária.

Estas descobertas mostram como o projeto poderia contribuir na análise destes casos, trazendo a possibilidade de maior rapidez na decisão, benefício defendido pelos estudiosos que apóiam o uso da jurisprudência.

6. CONCLUSÕES

Este trabalho teve como objetivo um exemplo de utilização de dados não estruturados visando a obtenção de uma maior gama de informações para estatísticas judiciárias.

O trabalho traz colaborações ao atual cenário de descoberta de conhecimento em novos tipos de formatos e na integração destes com um banco de dados convencional. Podemos citar como grandes contribuições: a análise e criação de técnicas para a busca de informações na estrutura do acórdão, uma extensão à ferramenta Kettle para realização desta busca e o desenvolvimento de uma estrutura para a integração entre os dados não estruturados e os convencionais estruturados.

As dificuldades encontradas quanto à busca de informações nos dados não estruturados já eram esperadas devido à falta de padrão deste tipo de dado. Também foi uma complicação ao projeto achar processos semelhantes. Ainda assim, foi possível encontrar padrões de decisão conforme esperado no início do projeto. Sendo assim, os objetivos iniciais foram alcançados.

6.1. Trabalhos Futuros

Outros estudos podem ser realizados a partir deste projeto, visando a melhora da integração e do aproveitamento da informação dos acórdãos.

O principal seria uma melhor forma de identificação da decisão dos desembargadores. Poderia ser considerado aqui o uso de ontologias. Este trabalho, apesar de bastante extenso, poderia trazer grandes contribuições ao tratamento de dados não estruturados.

Outros trabalhos semelhantes à este poderiam ser executados também visando o primeiro grau do sistema judiciário. Diversas informações poderiam ser observadas nesta instância, desde que se tivesse, com os dados não estruturados, o cuidado já citado na extração da sua informação.

7. REFERÊNCIAS BIBLIOGRÁFICAS

CASTERS, Matt; BOUMAN, Roland; DONGEN, Jos van. **Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration**, 2010.

DE PLÁCIDO E SILVA. **Vocabulário Jurídico Conciso**, 2009.

DEWSON, Robin. **Beginning SQL Server 2008 for Developers: From Novice to Professional**, 2008

DINIZ, Maria Helena. **Compêndio de Introdução à Ciência do Direito**, 1993.

IBM CORPORATION; **Designing and Modeling Databases**, Disponível em: <http://publib.boulder.ibm.com/> Acesso em: 2 maio. 2011

INMON, William. **Building the Data Warehouse**, 2002.

INMON, William. **Unstructured Applications: Unlocking the Potential**. Disponível em: <<http://www.inmoncif.com>>. Acesso em: 25 janeiro. 2011.

INMON, William; STRAUSS, Derek; NEUSHLOSS, Genia. **DW 2.0 The Architecture for the Next Generation of Data Warehousing**, 2007.

JUKIC, Nenad. **Data Modeling Strategies and Alternatives for Data Warehousing Projects**, 2006.

KIMBALL, Ralph. **The Data Warehouse Lifecycle Toolkit**, 1998.

KIMBALL, Ralph. **The Data Warehouse ETL Toolkit**, 2004.

KIETZMANN, Luís Felipe de Freitas. **Da uniformização de jurisprudência no direito brasileiro**. Disponível em: <<http://jus.uol.com.br/revista/texto/8701>>. Acesso em: 25 abril. 2011.

LEITE, Gisele. **Considerações sobre os Embargos de Declaração na sistemática recursal brasileiro** - Disponível em: <<http://www.clubjus.com.br/>>. Acesso em: 25 abril. 2011.

NETTO, Ernesto Junior Silveira. **A Influência da Jurisprudência no Direito Brasileiro. Artigo**. Disponível em: <<http://www.artigonal.com/>>. Acesso em: 20 abril. 2011.

PENTAHO, PENTAHO CORPORATION, Kettle. Disponível em:
<<http://kettle.pentaho.org/>> Acesso em: 10 maio. 2011

ROLDÁN, Maria Carina. **Pentaho 3.2 Data Integration: Beginner's Guide**, 2010.

STRECK, Lenio Luiz. **Súmulas no Direito Brasileiro: Eficácia, Poder e Função**.
Porto Alegre: Livraria do Advogado, 1998.

DW 2.0: Uma Forma de Tratamento Para Dados Não Estruturados em Acórdãos

ADAM VARGAS DE LIMA¹

¹UFSC – Universidade Federal de Santa Catarina
INE – Departamento de Informática e Estatística
Cx.P. 476 – CEP 88040-900 Florianópolis (SC)
{adam}@inf.ufsc.br

Resumo: Este trabalho de conclusão de curso tem como tema o tratamento e integração de dados não estruturados, assunto que vem ganhando crescente importância e que ainda se mostra bastante desafiador. Este projeto tem como objetivo a obtenção de uma estrutura eficiente para utilização em documentos jurídicos. O tema será primeiramente apresentado de forma teórica, buscando abranger as definições dos termos, processo de desenvolvimento, origens, vantagens de sua utilização e dificuldades encontradas no seu uso. Na parte teórica do trabalho, também serão apresentados os documentos a serem utilizados, os Acórdãos. Na parte prática do trabalho será apresentado, em um primeiro momento, o planejamento de todo o trabalho, passando pela análise dos documentos e técnicas para integração dos diferentes tipos de dados. A seguir, serão relatadas a execução dos passos previamente planejados e as dificuldades encontradas. Por fim, serão mostrados os resultados encontrados com a análise dos dados e comentada a possibilidade de trabalhos futuros na área.

Palavras-Chave: data warehouse, dados não estruturados, tomada de decisão.

1. Introdução

O sistema judiciário brasileiro possui hoje um grande banco de dados, o qual conta com o apoio de um sistema de Data Warehouse para geração de estatísticas e relatórios. Este DW é separado por instâncias do judiciário, as quais explicaremos melhor no capítulo sobre o Sistema Judiciário, e por estados da federação. Seu principal objetivo é prover informações sobre os processos.

Este sistema, assim como a quase totalidade dos sistemas de apoio à decisão atuais, foi desenvolvido visando o tratamento e um melhor armazenamento dos dados estruturados, estes obtidos diretamente de um banco de dados relacional. Os dados não estruturados não foram levados em conta neste desenvolvimento inicial. O motivo mais provável da sua não consideração é de que a ideia da sua utilização é recente, já que estes arquivos não foram previstos nas primeiras versões de Data Warehouse.

Seu uso é também consideravelmente mais complexo se comparado aos tradicionais dados estruturados, tanto na sua extração quanto na integração ao DW. Neste trabalho, utilizaremos estes dados, obtidos a partir dos acórdãos, para criarmos estatísticas sobre um conceito jurídico que vem ganhando cada vez mais força no nosso Sistema Judiciário: a jurisprudência.

1.1 Motivação

Durante muito tempo os desenvolvedores de sistemas de Data Warehouse mantiveram o foco apenas nos dados estruturados do banco de dados transacional. Os dados não estruturados recebiam pouca ou nenhuma atenção, fato este ocorrido em grande parte devido à falta de percepção sobre a quantidade de informação que poderia ser obtida utilizando-os. A importância destes dados foi notada diretamente na prática das empresas. Muito se perdia sem o armazenamento de dados encontrados em

arquivos como e-mails e notas de texto. O que já vinha sendo notado na prática, Bill Inmon expôs no livro DW 2.0 TM - Architecture for the Next Generation of Data Warehousing. Este novo paradigma vem ao encontro das necessidades observadas neste trabalho, quem tem como objetivo um eficiente tratamento dos acórdãos utilizando técnicas e conceitos criados por Inmon.

Ao extrairmos e estruturarmos a informação da lei usada nas decisões judiciais, esta hoje só encontrada nos arquivos não estruturados, poderemos descobrir padrões extremamente úteis, que poderão auxiliar no que tange ao uso desta fonte do direito. Visto que, em poder da lei utilizada, tornaremos possível a análise do que motivou o magistrado a tomar determinada decisão, a partir disto poderemos inferir se estas mesmas motivações foram usadas em outros casos parecidos.

Teremos então dados suficientes para a geração de estatísticas sobre o uso da jurisprudência.

Acreditamos que este projeto seja de grande valor para o aumento do uso e maior entendimento dos benefícios que podem ser adquiridos com estes dados.

- **Sistema Jurídico Brasileiro**

Para entendermos o motivo deste trabalho precisamos conhecer alguns pontos que envolvem um processo e suas diferentes fases, além da estrutura do acórdão e do conceito de jurisprudência.

- **O Processo**

Primeiramente, vamos conhecer de uma forma extremamente simplificada, apenas para entendimento do projeto, como é o andamento de um processo.

Na figura 1, vemos o percurso que um processo pode percorrer no Sistema Judiciário Estadual.

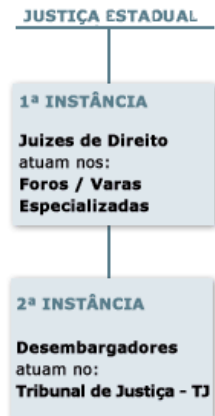


Figura 24 - Instâncias da Justiça Estadual
Fonte: elaborado pelo autor

Quando da entrada de um processo, primeiramente ele ingressa na primeira instância, ou primeiro grau. Em seguida é feita a distribuição do processo para algum magistrado da vara em questão, e então, após o julgamento, o juiz responsável por este processo dá a sua sentença.

Se uma das partes não concorda com a sentença dada em primeira instância, recorre à justiça de segundo grau para obter uma nova decisão.

Nesta segunda instância, o processo será analisado por um conjunto de magistrados, denominados desembargadores. O documento desta nova decisão é chamado de acórdão. Este documento é o foco do nosso trabalho.

- **Acórdão**

Conforme o Vocabulário Jurídico Conciso (DE PLÁCIDO E SILVA), um acórdão é “a resolução ou decisão tomada coletivamente pelos tribunais”. O que o diferencia da sentença, é, então, o seu caráter de decisão por um órgão colegiado, ao contrário da sentença que é emanada de um órgão monocrático. Simplificando, o acórdão é a decisão tomada por um grupo de desembargadores, os juízes dos Tribunais de Justiça dos Estados, proferido por um tribunal de justiça de segundo grau, que pode manter ou alterar a sentença do juiz de primeiro grau.

A estrutura geral de um acórdão e suas definições conforme De Plácido e Silva costuma ser:

- **Dados gerais**

Informações iniciais do processo em questão, divididas em Número de identificação da apelação, Apelante e Apelado.

- **Número de identificação da apelação.**

A apelação designa um dos recursos de que se pode utilizar a pessoa prejudicada pela sentença, a fim de que, subindo a ação à superior instância, e, conhecendo esta de seu mérito, pronuncie uma nova sentença, confirmando ou modificando, a que se proferiu na jurisdição de grau inferior.

- **Apelante.**

Diz-se da parte litigante, ou do terceiro prejudicado, que intentou o recurso de apelação sobre sentença, que lhe tenha causado gravame ou provocado prejuízo.

- **Apelado.**

Termo que designa a pessoa, que teve sentença favorável, de que se apelou.

- **Ementa.**

O resumo que se faz dos princípios expostos em uma sentença ou em um acórdão, ou o resumo do que se contém numa norma.

- **Relatório.**

Designa a exposição ou a narração, escrita ou verbal, acerca de um fato ou de vários fatos, com a discriminação de todos os seus aspectos ou elementos.

- **Voto.**

Manifestação da vontade, ou a opinião manifestada, pelo membro de uma corporação, ou de uma assembleia, acerca de certos fatos e mediante sistema, ou forma, preestabelecida. No acórdão, o magistrado explica o motivo de seu voto, baseando-se na constituição, para justificar sua decisão.

- **Decisão.**

Solução que é dada a uma questão ou controvérsia, pondo fim a ela, por meio de sentença, despacho ou interlocutória, e criando uma nova composição entre as partes contendoras ou litigantes. É, assim, o resultado de um pleito, quando é tida num sentido mais estrito, ou a mera deliberação a respeito de um ato ou de qualquer pedido que se faz no processo, numa aceção mais ampla.

- **Jurisprudência**

A definição de jurisprudência varia muito na linguagem técnica jurídica conforme o autor. Segundo Diniz (1993, p.290), jurisprudência é o conjunto de decisões uniformes e constantes dos tribunais, resultante da aplicação de normas a casos semelhantes, constituindo uma norma geral aplicável a todas as hipóteses similares e idênticas. É o conjunto de normas emanadas dos juízes em sua atividade jurisdicional. Para Miguel Reale (1998, p. 167), ela significa "a forma de revelação do direito que se processa através do exercício da jurisdição, em virtude de uma sucessão harmônica de decisões dos tribunais".

Interpretando estes autores, concluímos então, que o conceito de jurisprudência seria a consideração de mesmas ou semelhantes interpretações à casos análogos, utilizando como base para isso o registro de outras decisões anteriores e adaptando-os conforme as particularidades do processo em questão.

Com o maior uso dessa técnica, vantagens poderiam ser adquiridas. De acordo com Ernesto Junior Silveira Netto, "A jurisprudência evitaria que uma questão doutrinária ficasse eternamente aberta e desse margem a novas demandas, portanto diminuiria os litígios, reduziria os inconvenientes das incertezas do Direito, porque faria saber qual seria o resultado das controvérsias. Uma das maiores causas de queixas do sistemas judiciário é a lentidão, a jurisprudência viria em socorro desta demanda, possibilitando uma maior rapidez nas decisões uma vez que fornece subsídios valiosos ao magistrado." A principal vantagem do seu uso seria a segunda citada, a de uma maior rapidez no nosso sistema judiciário, como também afirma Luís Felipe de Freitas Kietzmann "Tem-se reconhecido cada vez mais a importância da jurisprudência no ordenamento jurídico pátrio, mormente quando se discute alternativas para desembaraçar o Poder Judiciário."

- **Data Warehouse**

Primeiramente, os mecanismos de armazenamento de dados eram simples, e não se preocupavam com nada além do puro armazenamento. Não tinham o intuito de buscar informações ou de apresentar dados de uma maneira que pudessem trazer novos tipos de estratégia às empresas. E de certa forma, isto nem era possível, já que cada dado custava muito quando armazenado em cartões perfurados. Com a introdução das fitas magnéticas essa primeira barreira foi vencida, o armazenamento passou a ser mais barato.

Mas ainda assim a leitura era difícil. Foi então que surgiu o armazenamento em disco, facilitando a última barreira, o acesso aos dados. A revolução foi grande, poderíamos agora guardar dados de forma relativamente barata e acessá-los, apagá-los ou reescrevê-los quando quiséssemos. O que acarretou no surgimento de uma massa de dados nunca antes vista, e, posteriormente, em um grande problema.

Dados com diferentes arquiteturas, em lugares diferentes, por vezes replicados e sem um padrão definido para o mesmo tipo. Tudo isso dificultava muito a obtenção de qualquer informação estratégica.

Foi nesse ambiente de caos que surgiu a idéia do processo de Data Warehousing, que basicamente era reunir esses dados, de forma que fosse desfeita toda essa bagunça em que haviam se transformado, a fim de gerar um suporte gerencial e estratégico às empresas.

O DW contém dados históricos integrados e granulares. Este é o seu grande segredo, esta integração possibilita às empresas uma capacidade de enxergar seu ambiente como um todo. Informações vindas de diversos lugares diferentes são apresentadas ao cliente como tendo uma fonte em comum.

Mas a implementação de um DW não é fácil. É preciso lidar com diversas dificuldades, como a integração dos dados vindos de diferentes fontes, o volume de dados, que por serem históricos e não-voláteis tendem a crescer exponencialmente e a diferença de desenvolvimento. Enquanto outros sistemas são construídos de uma vez só, um DW é construído em várias partes, tanto por seu tamanho, que tende a ser muito grande, quanto por seus requisitos, que tendem a mudar conforme o andamento do projeto. Assim, o DW acaba sendo um processo caro, complicado e de alto risco. O que faz com que as empresas pensem duas vezes antes de executá-lo.

Alguns conceitos da arquitetura e desenvolvimento de um DW serão explicados a seguir. Primeiramente veremos os conceitos de dimensão e fato, após isso os dois esquemas no qual um Data Warehouse pode ser modelado, o que é um Data Mart e, por último, seus métodos de desenvolvimento.

- **Dimensões**

Uma Dimensão é uma tabela que armazena aspectos textuais de cada elemento parte do processo.

As tabelas dimensionais contêm vários atributos que descrevem em detalhes todas as características que possam definir e serem úteis para futuras pesquisas no Data Warehouse. Podemos usar como exemplo uma dimensão para os Filmes de uma locadora. Nesta dimensão seriam armazenadas características dos filmes como Nome, Produtora, Atores, etc.

- **Tabelas Fato**

Enquanto as dimensões são tabelas textuais que buscam a descrição dos elementos, as tabelas fato tem como objetivo a análise quantitativa. Tabelas fato contem chaves estrangeiras para as dimensões e sua função é de medição. Por exemplo: número de locações de determinado gênero de filme.

- **Esquema Estrela**

Este tipo de modelagem é bastante simples e basicamente é composto de uma única tabela fato ligada a diversas dimensões. A simplicidade deste sistema é justamente uma de suas vantagens, já que proporciona facilidades ao usuário, o qual criará consultas com relacionamentos de baixa complexidade.

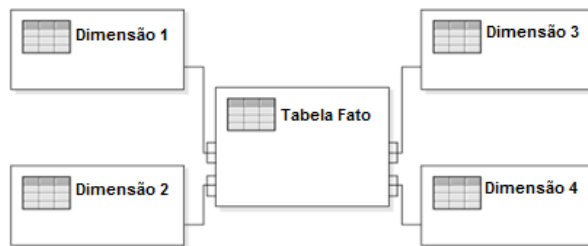


Figura 25 - Exemplo de esquema estrela

- **Esquema Floco de Neve**

Mais complexo que o esquema estrela, nesta modelagem as dimensões não necessariamente apontam para uma tabela fato.

Contrariamente ao esquema explicado anteriormente, busca a normalização das tabelas, com diversas dimensões podendo ser usadas para um tipo de informação. Para um Produto de uma companhia por exemplo. Suas consultas geralmente são mais

complicadas do que o esquema explicado anteriormente e primam mais pela rapidez na busca dos dados.

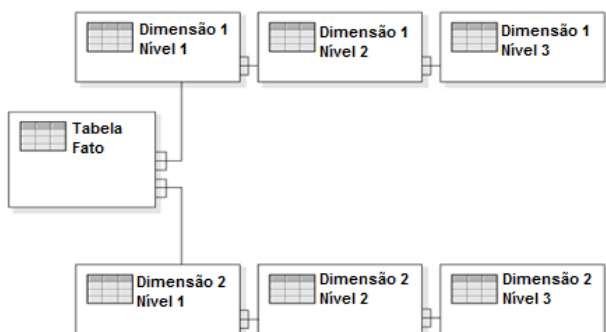


Figura 26 – Exemplo de esquema floco de neve

- **Data Mart**

Data Mart é um subconjunto de um Data Warehouse. São geralmente desenvolvidos a partir das necessidades de uma área da empresa ou do negócio. Como veremos a seguir, podem derivar de um Data Warehouse previamente construído, ou ter como fonte de dados a base transacional e, então, seu conjunto formar o DW. Por exemplo, em um tribunal de justiça, poderíamos ter um Data Warehouse de todas as informações que envolvem o Processo, e diversos data marts especificando uma informação sobre esta área de negócio, como Tempo de Andamento do Processo, Movimentação do Processo, etc.

- **Método de desenvolvimento Top-down**

Este método de desenvolvimento, defendido por Bill Inmon (2002), propõe que seja desenvolvido primeiramente um grande sistema de DW aonde serão concentradas todas as informações da organização relevantes à tomada de decisão independentemente da área. Para uma definição simples, podemos citar Jukic(2006) que cita o Data Warehouse como sendo “uma fonte de dados para os novos Data Marts”. A partir desta fonte de dados central seriam criados os diversos Data Marts necessários. É um modelo consistente quanto à mudanças no negócio. Sua principal desvantagem é o custo para desenvolvimento do grande projeto inicial, tanto na questão de tempo de implementação quanto de pessoas para isto.

- **Método de desenvolvimento Bottom-up**

Ao contrário de Inmon, Ralph Kimball (1998) defende que o projeto seja executado a partir de projetos pequenos, os Data Marts, para então, o conjunto destes compor o Data Warehouse da organização.

Neste método de desenvolvimento, os dados para o Data Mart vem diretamente da base transacional, sendo esta a fonte de dados. Como os Data Marts são independentes, assim que um deles é finalizado a organização já pode usa-lo. Como exemplo, se o Data Mart “Vendas” foi desenvolvido, e o “Marketing” ainda está sendo criado, a empresa já pode criar estatísticas a partir do primeiro. Esta é uma grande vantagem, fazendo com que a companhia obtenha resultados rapidamente, mesmo que ainda não tenha um Data Warehouse completo.

- **DW 2.0**

Evolução do DW. É assim que podemos tratar o DW 2.0 descrito por Inmon no livro DW 2.0 TM - Architecture for the Next Generation of Data Warehousing. Bill Inmon sendo um dos elaboradores do conceito inicial de Data Warehouse, com o tempo identificou alguns problemas deste primeiro trabalho. Alguns observados, através da experiência após anos de implementação no mercado profissional, outros causados pela falta de registro do conceito. Para entendermos essa nova abordagem, se torna necessário primeiramente uma visão geral sobre as limitações da primeira ideia de um armazém de dados. Surgido academicamente nos anos 80 e com

forte intuito de ser usado comercialmente, após ser colocado em prática, foram observados novos problemas e requisitos a serem sanados pelos sistemas de apoio à decisão. Como descreve Inmon, “muitas forças moldaram a evolução da arquitetura de informação a este mais alto nível – DW 2.0”.

Entre elas podemos citar a demanda por diferentes tecnologias. A evolução já comentada relacionada ao armazenamento de dados trouxe ao usuário uma forma de interação e uma quantidade de informação cada vez maior. Conforme citado anteriormente, em pouco tempo a computação passou de cartões perfurados a um armazenamento e processamento de dados que tornaram possíveis a criação de telas com uma incrível diversidade de informações, apresentadas das mais diferentes maneiras.

Devido a esta quantidade de informação, passou a ser preocupante e um objeto de estudo o tempo e a facilidade de acesso à elas. Notou-se então que um dado de cinco anos atrás tem um padrão de acesso totalmente diferente de um dado do mês corrente, o que levou a proposta do ciclo de vida de um dado.

De acordo com a proposta de ciclo de vida do dado no DW 2.0, um dado é tratado de diferentes formas conforme sua necessidade de acesso. Necessidade esta que é baseada na “idade” do dado. De acordo com essa teoria, quanto mais antiga a informação, menor é a necessidade de performance na sua busca. São usados então, para dados mais antigos, formas de armazenamento que privilegiem o tamanho da base ao invés da rapidez na sua recuperação. O que faz com que o DW tenha um custo muito menor e uma agilidade muito maior no acesso aos dados.

Outro problema surgido com a experiência no mercado foi a necessidade de inclusão de dados não estruturados. Muitas informações podem ser adquiridas de e-mail, notas, relatórios, etc. Nada disso havia sido considerado na primeira abordagem. O reconhecimento da importância desses requisitos levou ao DW 2.0, e é a implementação desta última característica o motivo deste projeto.

• **Dados não estruturados**

Desde o desenvolvimento dos primeiros Data Warehouses, as decisões foram sempre baseadas em dados obtidos de tecnologias estruturadas, negligenciando assim as formas de informação não estruturadas, como e-mail, notas e relatórios. Inmon comenta no artigo *Unstructured Applications: Unlocking the Potential*, “Cruzando o vazio entre

dados não estruturados e dados estruturados, uma nova perspectiva de dados é possível. Fazendo uma ponte entre o vazio entre os dois tipos de ambiente é possível combinar texto e dados numéricos. Esta habilidade possibilita novos tipos de sistemas totalmente novos a serem construídos”.

Um exemplo do que Inmon afirma é a rapidez de informação que um e-mail pode trazer. Este tipo de dado nos mostra informação imediata de um contato com o cliente. Se antes só se sabia dos dados da compra de uma determinada pessoa, agora se pode saber o que essa pessoa pensa e fala. Isso tem um valor incomensurável quando pensamos em quão rápido os mercados mudam e quão grande é o esforço das empresas para manter o seu cliente. Ficou muito mais fácil traçar todo o perfil de um indivíduo ao tornar tangíveis informações que até então eram inalcançáveis. Aqui também podemos dar um exemplo do armazenamento desses dados. Ao procurar por dados estruturados de determinado cliente no banco, o analista pode ou receber uma indicação do lugar onde se encontram os e-mails de comunicação deste com a empresa, ou trazer o e-mail junto com os outros dados, o que de fato traria facilidades na análise, mas necessitaria de um poder de processamento muito maior.

Empresas de vanguarda e executivos mais atentos às necessidades de suas companhias já observaram o que Inmon argumenta e isso vem mudando rapidamente. As empresas cada vez mais se dão conta de que não podem desconsiderar esses dados e a informação estratégica contida neles. Como afirmou Erik Moller, responsável de Information Management da HP Software à CXO, “Os dados não estruturados são quase um tabu para as empresas, que terão sérias dificuldades em compreender totalmente a sua informação de negócio se continuarem a ignorar esta questão”. Ainda de acordo com Moller: “os CIO prevêm uma redução significativa na informação não estruturada das empresas ao longo dos próximos três anos. Mas verificamos que a maioria das empresas não tem a noção da quantidade de dados não organizados que existe no seio das suas organizações”.

Ele sugere que as companhias atentem para isso e tentem o mais rápido possível automatizar a gestão desses documentos, já que, segundo uma pesquisa da própria HP, 70% dos dados ainda são perdidos por serem não estruturados e não receberem a devida atenção. Uma explicação para isso é que a princípio pode parecer algo complicado e até mesmo desnecessário, mas vale lembrar que, conforme a análise de Martin Atherton, analista da consultora

Freeform Dynamics, "Podemos debater incessantemente a questão da gestão de informação, mas o que é importante é que as empresas percebam que os recursos de informação ao seu dispor podem ajudá-las a tomar as suas decisões de uma forma mais rápida e eficaz".

Assim, é uma forma valiosíssima a mais de informação, e, da mesma maneira que no início um DW era algo considerado muito dispendioso e talvez dependente de um esforço exagerado, hoje é uma unanimidade entre empresas preocupadas com a estratégia empresarial. O DW 2.0 tende a se mostrar com o tempo tão bem-sucedido quanto o seu predecessor.

- **Dificuldades**

Graças a sua imprevisibilidade já comentada, alguns cuidados devem ser tomados quando lidamos com estes dados. Inmon cita os seguintes problemas: conversas sem importância, terminologia e texto específico ou geral demais. Como exemplos de conversa sem importância, podemos citar e-mails pessoais trocados entre funcionários. Para uma empresa, de nada importa saber de assuntos entre um funcionário e sua esposa. Grande parte dos e-mails trata disso. Atrapalham muito e devem ser eliminados do nosso procedimento.

O problema da terminologia também traz dificuldades. Pessoas de todo o tipo influenciam os dados não estruturados. Assim, é normal que conforme a idade, classe social, cultural, entre outras diferenças, as pessoas escrevam de maneiras diversas. Uma mesma palavra pode significar coisas diferentes nesse contexto. Outro problema é que, como o autor do dado provavelmente não tem conhecimento de que seu texto será usado em um DW, não existe a preocupação em utilizar determinado conjunto de palavras. Sendo necessário então, que um analista prepare os documentos de uma forma que seja interpretado pelo computador como deve ser, evitando dados equivocados. Esse processo não é fácil e é uma importante parte do DW.

Deve-se então, buscar a normalização do texto, a ser feito de duas formas: específica e genérica. Um exemplo desse conceito, conforme Inmon: "O Tarsus foi submetido à pressão e desarticulado". Assim seria a versão específica de um texto. Mas se alguém procurar por "osso quebrado", não trará esse registro como resultado. Procurando por osso não trará Tarsus e quebrado não encontrará desarticulado. Mas se o texto se encontrar nas duas formas, específico e

genérico, o termo Tarsus será reconhecido como osso e desarticulado será reconhecido como quebrado. Assim, a partir da frase em questão, dois conjuntos de dados seriam criados: tarsus/osso e desarticulado/quebrado. Este é o segundo grande passo para o tratamento de dados não estruturados, e, sem ele a análise pode perder o sentido, fazendo com que o DW 2.0 se torne um fracasso, não trazendo toda a informação que poderia.

- **Visão geral dos dados não estruturados no DW 2.0**

Para a carga dos dados não estruturados no banco de dados a fim de torná-los úteis para o DW, são necessários alguns processos. Primeiramente, um processo de ETL, que, como previsível, é diferente do processo de extração e tratamento dos dados estruturados. Este processo será detalhado futuramente, e algumas de suas atividades são:

- Padronização
- Remoção de palavras sem utilidade
- Tratamento de ortografia alternativa
- Suporte às buscas
- Uso de Ontologias externas ou internas

Após a realização deste processo, o conteúdo da fonte não estruturada está pronto para ser inserido no banco. Outros tipos de dado fazem parte da estrutura proposta por Inmon. São eles:

- Taxonomias internas e externas: Uma taxonomia é uma lista de palavras nas quais existe alguma relação entre as palavras. O ambiente textual não estruturado inclui taxonomias que foram criadas internamente (algumas vezes chamadas de “temas”) e taxonomias externas que podem vir praticamente de qualquer lugar.
- Texto editado e capturado: Texto editado e capturado é o texto que passou através do processo ETL não estruturado e foi colocado em uma base de dados relacional padrão.
- Links: Links são os dados que amarram os dados não estruturados aos dados estruturados.
- Apontadores simples: Ocasionalmente os dados não estruturados irão permanecer em outro ambiente e apenas índices de referência para isso serão trazidos para o setor interativo de dados não estruturados do DW.

Outra característica deste tipo de dado é que, como mostrado na figura 4, geralmente apenas duas atividades são ligadas a eles: a carga na base e o acesso.

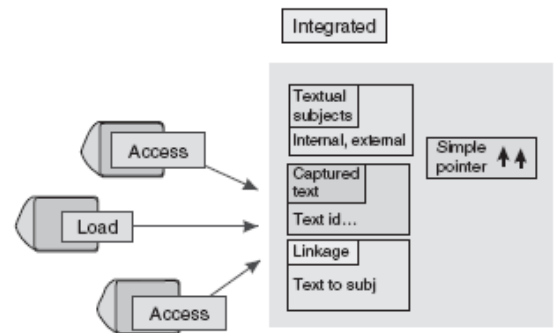


Figura 4 - Acesso aos dados não estruturados na base

Não é comum e nem recomendada a atualização dos dados textuais não estruturados. De acordo com Inmon, depois que uma descrição textual do trabalho é realizada, se mudanças necessitam ser feitas ela é completamente reescrita. Dados textuais incrementais ou parcialmente atualizados simplesmente não são um reflexo da realidade.

- **Envolvimento do Usuário**

Se existem dados próximos ao usuário, estes dados são os textuais não estruturados. Dados textuais não estruturados estão presentes no dia-a-dia do usuário final. Assim, o usuário final é altamente envolvido no processo de inclusão de textos não estruturados no DW 2.0 (INMON; STRAUSS; NEUSHLOSS, 2007). O usuário final está presente durante quase todo o processo. Isto ocorre pela necessidade que o tratamento de dados não estruturados tem de tratar os dados com conhecimento de negócio. Como exemplo, temos sua participação na especificação de palavras sem utilidade e terminologia.

Geralmente o usuário do negócio tem somente um envolvimento passivo na modelagem dos aspectos estruturados no DW 2.0. Mas o contrário é verdadeiro para os aspectos não estruturados. Por exemplo, o usuário é profundamente envolvido nas especificações da ETL dos dados não estruturados (INMON; STRAUSS; NEUSHLOSS, 2007).

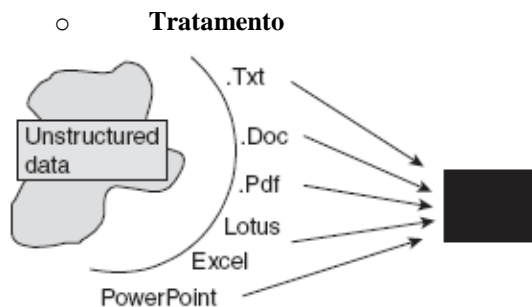


Figura 5 - Tratamento dos dados não estruturados

O primeiro passo no processo de preparação dos dados não estruturados para o processo de análise é a “leitura” do texto pelo sistema. O texto vem em uma considerável variedade de formatos. Os formatos podem precisar ser lidos como entrada. Depois que a fonte original foi lida, o próximo passo é preparar os dados para entrar na base. Esta preparação textual envolve um processo para que os dados possam ser inseridos na base.

Existem várias boas razões do motivo deste processamento, entre elas:

- Dados não estruturados precisam ser ajustados em um formato relacional;
- Dados não estruturados precisam ser integrados para que um processo de análise bem sucedido possa ser feito. Se um texto

sem análise é simplesmente “jogado” na base, essa análise pode perder o sentido.

- **ETL**

Uma importante decisão deve ser feita no momento anterior à integração dos dados: em que ambiente fazer o tratamento? Pode ser feito no estruturado ou não estruturado. Para que seja feito no lado estruturado é necessária a adaptação dos dados. Pode parecer muito trabalhoso, mas esta adaptação possibilita o uso de tecnologias de análise padrão. Além disso, como o mundo da análise de dados sempre foi moldado em torno dos dados estruturados, muito já foi gasto em treinamento de usuários e equipes técnicas no ramo de inteligência de negócios visando estes dados. Banco de dados, ETL e processo estatístico já foram desenvolvidos com este propósito. Não faz sentido então, desprezar toda essa tecnologia desenvolvida buscando um novo desenvolvimento caro e trabalhoso. Diante de tudo isso, a escolha se torna simples: o ambiente estruturado é o ideal.

O processo de “integração” textual antes da alocação na base de dados passa por diferentes etapas. Segundo Inmon, uma série de passos é necessária na preparação do texto para incorporação na base de dados e posterior análise no DW 2.0. São elas:

- Padronização
- Remoção de palavras sem utilidade
- Substituição ou concatenação de sinônimos
- Detalhamento de expressões
- Criação de temas
- Uso de Glossários ou Taxonomias externas
- Radical
- Ortografia alternativa
- Internacionalização
- Suporte a busca direta e indireta

- **Planejamento**

Desnecessário falar sobre a importância do planejamento antes da execução de qualquer sistema de informação. Principalmente quando o assunto é novo e sem muitos exemplos, como a extração de informações a partir de documentos não estruturados para um Data Warehouse. O planejamento, segundo Reynolds capacita administradores a direcionar os esforços e recursos da organização para o alcance de objetivos específicos.

É então um passo que evita o desperdício de tempo e recursos no desenvolvimento de um sistema que

depois se mostrará insuficiente para o atendimento dos requisitos do negócio.

○ **Proposta do Trabalho**

Conforme ilustrado na figura 6, a visão geral do trabalho será:

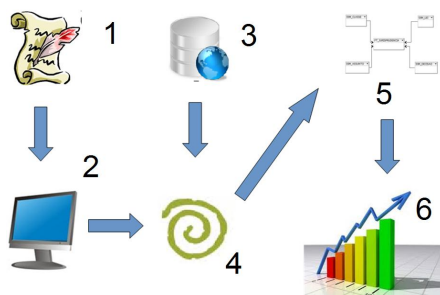


Figura 6 - Estrutura geral do projeto

Fonte: elaborado pelo autor

1. Estudo de quais partes do acórdão nos serão úteis
2. Desenvolvimento do plugin para extração de dados a partir desta análise
3. Preparação dos campos estruturados a serem integrados.
4. Extração dos dados não estruturados e integração com os dados vindos das tabelas estruturadas, utilizando o Kettle.
5. Desenvolvimento dos Data Marts.
6. Informações obtidas a partir da extração da informação dos acórdãos.

Segue um maior detalhamento de cada uma destas etapas.

● **Análise do Acórdão**

A análise do negócio é sempre uma das partes principais, senão a mais importante, de um Data Warehouse. O sucesso depende, em grande parte, da correta identificação dos problemas e de como resolvê-los através das ferramentas que temos em mãos. Nosso problema é a falta de dados para apoio da jurisprudência, e nossa fonte para a resolução desta carência é a extração de informações a partir do acórdão.

Conforme já mencionado, um acórdão tem a seguinte estrutura: dados do processo, ementa, relatório, voto e decisão. A parte de relatório do processo é, por vezes, bastante extensa, o que dificulta a consulta e análise. Justamente nesta parte se encontra um dado chave para o nosso projeto, a lei usada na decisão.

Tendo como ideia deste trabalho possibilitar o acesso às estatísticas do emprego da jurisprudência ou, até mesmo, facilitar o acesso direto à conclusão de determinado processo, o tratamento adequado para este documento é a obtenção de um resumo da decisão, trazendo o voto do magistrado, e em qual lei foi baseada esta sentença.

Hoje em dia já são armazenados no sistema a Classe e o Assunto do processo, responsáveis por definir o tipo do processo. Ao unirmos a isso a lei utilizada e a decisão tomada a partir dela, possibilitaremos a criação de estatísticas sobre a Jurisprudência.

● **ETL do documento não estruturado**

Depois de decidido, na análise do acórdão, quais os pontos necessários para que o documento se ajuste às necessidades do trabalho, foram definidos os passos da extração de documentos não estruturados proposta por Inmon que melhor se adaptam ao caso do tratamento do acórdão.

A partir desta decisão, foi desenvolvida uma extensão utilizando a linguagem de programação Java. Este futuramente será integrado ao Kettle, software para realização de Integração de dados já citado.

Com a definição de quais passos propostos na teoria do DW 2.0 poderiam ser usados neste trabalho, identificamos quais características deveriam ser desenvolvidas no plugin.

- Primeiramente a padronização, fazendo com que as palavras fiquem todas com a mesma fonte e com letras minúsculas, facilitando as buscas no texto.
- O segundo passo é a remoção de palavras e termos sem utilidade. Como a ideia é deixar o documento útil para ser usado no apoio à jurisprudência, na análise foi decidido por buscar a decisão do magistrado e a lei na qual essa decisão foi baseada.
- A seguir, a substituição de sinônimos, passo utilizado nas decisões presentes nos documentos, que podem usar diferentes palavras para um mesmo significado.

Buscando a integração com o DW atual, será necessário também o número do processo, de forma que, a partir deste código, conseguiremos ligar o acórdão ao seu processo já cadastrado na base de dados.

Sendo assim, durante o tratamento feito no acórdão, extrairemos o número do processo, a decisão tomada e a lei. A decisão e a lei se tornarão duas dimensões com chaves em colunas da nossa tabela e o número do processo foi usado para ligarmos estas informações ao seu respectivo processo.

- **Integração ao banco de dados estruturado**

Para o projeto de integração, é necessário o entendimento do padrão atual do DW.

Podemos, hoje, gerar diversas estatísticas baseadas em fases ou características do processo. Para isso, existem vários Data Marts, como: Andamento Médio dos Processos, Movimentação do Processo e Tempo Médio do Processo. No sistema atual, cada Data Mart possui somente as chaves de dimensões consideradas importantes para identificação de processos semelhantes e uma coluna quantitativa. Isso possibilita a criação de estatísticas.

Como exemplo, em uma tabela fato responsável por informar estatísticas sobre o tempo do processo, utilizaríamos chaves relativas ao tema do processo (Dim_Classe e Dim_Assunto), ao local onde o processo se encontra (Dim_Tribunal) e ao tempo (Dim_Tempo). A figura 9 mostra um exemplo de Data Mart do Data Warehouse atual, com todas as dimensões tendo como fonte, os dados estruturados.

Após análises efetuadas no documento, vimos que serão necessárias duas novas dimensões para o nosso projeto, visto que pretendemos criar estatísticas baseadas na Lei e na Decisão judicial. Baseado nesse esquema planejamos nossa tabela para os dados não estruturados.

A tabela fato Jurisprudencia terá três chaves comuns às demais tabelas semelhantes do banco, essas tem como fonte os dados estruturados, são elas: a Dimensão Classe, a Dimensão Assunto e a Dimensão Tempo. Junto a isso, a tabela terá também ligação com as duas novas dimensões, provenientes dos dados não estruturados, a Dimensão Lei e a Dimensão Decisão. Será possível então, através desta, verificar quantos processos de mesma classe e assunto foram decididos baseados na mesma lei e tiveram a mesma decisão. Detalharemos melhor as dimensões a seguir.

- **Dimensão Classe**

Dimensão já presente no banco de dados atual. Será integrada aos dados do acórdão pelo

código do processo. As classes são determinadas pelo Conselho Nacional de Justiça e, junto com o assunto, informam a natureza do processo.

- **Dimensão Assunto**

Assim como a Dimensão Classe, esta já existe no DW atual e sua origem são os dados estruturados. Indica do que se trata o processo.

Responsável no sistema por apontar processos similares para a geração de estatísticas.

- **Dimensão Lei**

Informação impossível de ser obtida atualmente a partir dos dados estruturados, será preenchida a partir da extração nos acórdãos.

Indicará quais leis foram utilizadas para a tomada de decisão. Junto com a decisão é a principal inovação deste trabalho e o que tornará possível o seu objetivo.

- **Dimensão Decisão**

Informará a decisão do órgão colegiado, conforme descrito no acórdão. Também preenchida a partir dos dados não estruturados. Indicará, junto com a dimensão lei, a resolução do processo e os motivos.

- **Análises**

Depois de tratados e indexados, os documentos estão prontos para as funcionalidades do DW. Nesta etapa serão geradas as estatísticas relativas ao uso da jurisprudência e poderemos também efetuar a busca em uma decisão de um processo específico.

- **Integração dos Dados**

Depois do planejamento, é o momento de testar e avaliar o sistema. Apresentaremos os resultados gerais, quanto às fontes de informação utilizadas, cargas de dados nas dimensões e outras tabelas e por fim analisaremos as informações obtidas.

- **Fontes de informação**

Depois do planejamento, é o momento de testar e avaliar o sistema. Neste capítulo apresentaremos os resultados gerais, quanto às fontes de informação utilizadas, cargas de dados nas dimensões e outras tabelas e por fim analisaremos as informações obtidas.

7.1.1. Acórdãos

Os acórdãos utilizados foram extraídos do DVD Jurisprudência Catarinense 2009, obtido no Tribunal de Justiça de Santa Catarina. Como não existe, no DVD, a possibilidade de extração dos arquivos, os mesmos foram copiados e salvos em arquivos de textos para que pudessem servir ao projeto. Foram utilizados diversos acórdãos de diferentes tipos, de modo que se pudesse analisar individualmente a correta integração e geração de estatísticas. Conforme citado no planejamento, as informações extraídas dos acórdãos foram o código do processo, a lei utilizada e a decisão final.

7.1.2. Dados estruturados

A fonte dos dados estruturados foram dados de teste do banco de dados já existente no Sistema Judiciário de Santa Catarina. Conforme definido no planejamento, os dados estruturados buscados foram a classe e o assunto de cada processo.

○ **Integração com o DW**

As duas fontes de dados citadas no capítulo anterior são o ponto inicial do processo. Após a entrada no processo de ETL, os acórdãos passam pelo tratamento para extração das informações necessárias. Este tratamento é feito através do plugin desenvolvido para o Kettle. Os campos da lei utilizada e da decisão tomada, após a extração feita com a extensão desenvolvida, formarão duas dimensões: a Dim_Lei e a Dim_Decisao. Depois de obtido o código do processo, através dele será realizada a integração com o banco estruturado. A partir deste código, será feita a junção para a busca da classe e do assunto do processo. Neste ponto do trabalho já teremos todas as informações necessárias para o preenchimento da tabela fato.

○ **Extensão Desenvolvida**

O plugin desenvolvido para o projeto foi criado visando o tratamento dos dados do Acórdão.

Seu processo é simples: após a entrada do documento no fluxo, realizada através do Kettle, é feita uma busca no documento, procurando por palavras chaves previamente definidas de acordo com a análise dos acórdãos. Estas palavras definirão as informações que buscamos. Por exemplo: ao encontrarmos “nego o provimento”, saberemos que a decisão foi de negar a Apelação.

Quando alguma das palavras é encontrada, seu significado no caso da decisão ou a própria palavra no caso da lei é guardada e colocada no fluxo da ETL. Como seu procedimento é totalmente interno e padrão, não existe uma tela de interface com o usuário.

Para um melhor entendimento, analisaremos o código desenvolvido. Primeiramente é necessária a explicação de que o software Kettle envia para o plugin cada linha do acórdão individualmente, por este motivo a análise foi então feita linha a linha.

O início do documento, aonde nos é informado o número do processo, tem o seguinte padrão:

Apelação nº: 2934856

Por aparecer no início do documento, o número da apelação é a primeira informação a ser buscada. Para isto, primeiramente conferimos se esta palavra existe no documento, através da função `lastIndexOf`, nativa da linguagem Java. Caso o conjunto de caracteres exista, esta função retornará “-1”. Após a confirmação da sua existência, buscaremos a informação do número do processo, o que será feito através da função `substring`, retornado os caracteres posteriores à “Apelação nº:”, ou seja, o

número do processo, que no nosso exemplo seria “2934856”, esta informação será então colocada no fluxo já apresentado e não será mais buscada nas próximas linhas recebidas pela nossa extensão.

A seguir, buscaremos nas próximas linhas recebidas o número da Lei. Em todos os acórdãos analisados foi observado o mesmo padrão nesta informação, o uso da expressão “Lei n.º” antes do número da lei utilizada. Seguindo o exemplo anteriormente citado, receberíamos a seguinte linha:

da Lei n.º 8.429/92, bem como ao pagamento de honorários

Um procedimento muito semelhante ao utilizado para a extração do número da apelação é utilizado, com a diferença de que como podemos ter outras informações após o número da lei, a função substring retorna o conjunto de caracteres desde o final de “Lei n.º” até o próximo espaço vazio.

Depois de já termos preenchido o número do processo e a lei utilizada, procuraremos nas próximas linhas por uma série de palavras que podem determinar a decisão do juiz e as enquadramos em um dos três tipos: provimento concedido, provimento negado ou provimento parcial.

Depois de recebidas estas informações, elas são colocadas em suas respectivas variáveis e inseridas no fluxo do Kettle para inserção nas tabelas.

- **Processo de ETL**

7.1.3. Dimensão Lei

Primeiramente vamos falar da Dimensão Lei. Denominada Dim_lei no nosso DW. Conforme mostrado na figura 7, é uma dimensão extremamente simples. Seus campos são o CHLEI, a chave da dimensão, e a descrição da lei. Como podemos ver na figura, o campo CHLEI é preenchido diretamente pelo banco de dados.

Dim_lei				
	Column Name	Condensed Type	Nullable	Identity
🔑	CHLEI	bigint	No	<input checked="" type="checkbox"/>
	deLei	varchar(MAX)	No	<input type="checkbox"/>
				<input type="checkbox"/>

Figura 7 - Dimensão Dim_lei

O plugin vai percorrer o documento procurando as informações já citadas. A seguir é excluída qualquer linha vazia que possa aparecer, etapa executada apenas por segurança e então encontrada a informação de Lei requerida.

7.1.4. Dimensão Decisão

A Dimensão Decisão é bastante similar à de Lei, tanto na sua estrutura quanto no seu tratamento.

Dim_decisao				
	Column Name	Condensed Type	Nullable	Identity
🔑	CHDECISAO	bigint	No	<input checked="" type="checkbox"/>
	deDecisao	varchar(MAX)	No	<input type="checkbox"/>
				<input type="checkbox"/>

Figura 8 - Dimensão Dim_decisao

A dimensão é formada por uma chave e pela descrição da decisão. Esta, como a descrição da lei, também será obtida a partir do tratamento do acórdão. Mas, enquanto na descrição relativa à Lei ela simplesmente é inserida na tabela, na descrição da decisão é realizado um tratamento pelo plugin. A dimensão receberá o significado da decisão, e não uma transcrição do Acórdão.

Este passo já foi explicado na seção sobre o desenvolvimento do Plugin. Esta adaptação tem por motivo a criação de estatísticas, de modo que o importante é termos o resultado da decisão, independente de como foi transcrita.

- **Jurisprudência**

A tabela fato Jurisprudencia será a fonte das estatísticas a serem geradas. Tem como colunas as chaves das dimensões e um campo totalizador.

Jurisprudencia			
Column Name	Condensed Type	Nullable	Identity
CHLEI	bigint	No	<input type="checkbox"/>
CHDECISAO	bigint	No	<input type="checkbox"/>
CHCLASSE	bigint	No	<input type="checkbox"/>
CHASSUNTO	bigint	No	<input type="checkbox"/>
totalProces...	int	No	<input type="checkbox"/>

Figura 9 - Tabela fato Jurisprudencia

Com os dados desta tabela poderemos ver, por exemplo, quantos processos foram negados do assunto Furto, da classe Habeas Corpus que usaram determinada lei para esta decisão.

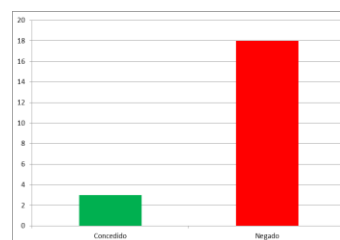
- **Resultados**

Após a análise dos acórdãos, o desenvolvimento do plugin, o estudo para integração com os dados estruturados e a criação de tabelas, foi executada a carga dos acórdãos a partir de dados reais obtidos no DVD Jurisprudência Catarinense 2009. Foram extraídos deste DVD cerca de 200 acórdãos, a fim de testar a integração com o DW atual e possibilitar o descobrimento de algum padrão nas decisões dos magistrados. Quanto ao preenchimento das dimensões, o processo obteve grande sucesso. A única observação a ser feita é a dificuldade em relação à identificação das palavras que identificam a decisão, visto que, diferentes termos são utilizados dependendo do relator do acórdão. O tratamento destes casos, previsto por Inmon e descrito na seção sobre tratamento dos dados não estruturados no item “Substituição de sinônimos”, foi considerado no desenvolvimento do plugin. Mesmo assim, como são expressões que mudam constantemente, será preciso um melhor tratamento quando a carga de acórdãos for consideravelmente maior.

A integração dos dados também foi bem sucedida. Com o número do código do processo as chaves de Classe e Assunto foram ligadas de maneira correta aos processos dos acórdãos, assim como as dimensões da lei utilizada e da decisão tomada. Na carga dos

acórdãos e na criação de estatísticas foi observado um problema. Os acórdãos variam muito de assunto e classe, e como a base destes documentos não tinha uma organização segundo o tipo de processo, foi extremamente complicado o acompanhamento de processos semelhantes.

Mesmo com estas dificuldades, foi possível a análise de processos de Habeas Corpus relativos à prisões pelo assunto tráfico de drogas. Através da análise da estatística desses processos notamos uma clara tendência quanto à jurisprudência deste caso específico, como podemos observar no gráfico da figura 10.



Podemos notar um claro padrão quanto às decisões. De cada 21 acórdãos do tipo citado solicitando a suspensão da prisão, 18 foram negados. Portanto, este caso é um exemplo de que o objetivo do projeto foi atingido, ou seja, a correta extração e integração de dados não estruturados possibilitando a descoberta de padrões nas decisões da justiça do segundo grau.

Outros padrões também foram descobertos. Em 2009 houve diversos casos em que os professores não receberam determinados benefícios por estarem de licença devido a motivos de saúde. Todos os acórdãos analisados consideraram isso ilegal e decidiram favoravelmente aos professores.

Além destes dois usos da jurisprudência na decisão dos magistrados, o mesmo padrão foi observado na decisão quanto à transferência de ações quando da compra de uma empresa de telefonia estatal por uma empresa privada.

Estas descobertas mostram como o projeto poderia contribuir na análise destes casos, trazendo a possibilidade de maior rapidez na decisão, benefício defendido pelos estudiosos que apóiam o uso da jurisprudência.

- **Conclusões**

Este trabalho teve como objetivo um exemplo de utilização de dados não estruturados visando a obtenção de uma maior gama de informações para estatísticas judiciais.

O trabalho traz colaborações ao atual cenário de descoberta de conhecimento em novos tipos de formatos e na integração destes com um banco de dados convencional. Podemos citar como grandes contribuições: a análise e criação de técnicas para a busca de informações na estrutura do acórdão, uma extensão à ferramenta Kettle para realização desta busca e o desenvolvimento de uma estrutura para a integração entre os dados não estruturados e os convencionais estruturados.

As dificuldades encontradas quanto à busca de informações nos dados não estruturados já eram esperadas devido à falta de padrão deste tipo de dado. Também foi uma complicação ao projeto achar processos semelhantes. Ainda assim, foi possível encontrar padrões de decisão conforme esperado no início do projeto. Sendo assim, os objetivos iniciais foram alcançados.

○ **Trabalhos Futuros**

Outros estudos podem ser realizados a partir deste projeto, visando a melhora da integração e do aproveitamento da informação dos acórdãos.

O principal seria uma melhor forma de identificação da decisão dos desembargadores. Poderia ser considerado aqui o uso de ontologias. Este trabalho, apesar de bastante extenso, poderia trazer grandes contribuições ao tratamento de dados não estruturados.

Outros trabalhos semelhantes à este poderiam ser executados também visando o primeiro grau do sistema judiciário. Diversas informações poderiam ser observadas nesta instância, desde que se tivesse, com os dados não estruturados, o cuidado já citado na extração da sua informação.

• **Referências**

CASTERS, Matt; BOUMAN, Roland; DONGEN, Jos van. **Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration**, 2010.

DE PLÁCIDO E SILVA. **Vocabulário Jurídico Conciso**, 2009.

DEWSON, Robin. **Beginning SQL Server 2008 for Developers: From Novice to Professional**, 2008

DINIZ, Maria Helena. **Compêndio de Introdução à Ciência do Direito**, 1993.

IBM CORPORATION; **Designing and Modeling Databases**, Disponível em: <http://publib.boulder.ibm.com/> Acesso em: 2 maio. 2011

INMON, William. **Building the Data Warehouse**, 2002.

INMON, William. **Unstructured Applications: Unlocking the Potential**. Disponível em: <<http://www.inmoncif.com>>. Acesso em: 25 janeiro. 2011.

INMON, William; STRAUSS, Derek; NEUSHLOSS, Genia. **DW 2.0 The Architecture for the Next Generation of Data Warehousing**, 2007.

JUKIC, Nenad. **Data Modeling Strategies and Alternatives for Data Warehousing Projects**, 2006.

KIMBALL, Ralph. **The Data Warehouse Lifecycle Toolkit**, 1998.

KIMBALL, Ralph. **The Data Warehouse ETL Toolkit**, 2004.

KIETZMANN, Luís Felipe de Freitas. **Da uniformização de jurisprudência no direito brasileiro**. Disponível em: <<http://jus.uol.com.br/revista/texto/8701>>. Acesso em: 25 abril. 2011.

LEITE, Gisele. **Considerações sobre os Embargos de Declaração na sistemática recursal brasileiro** - Disponível em: <<http://www.clubjus.com.br/>>. Acesso em: 25 abril. 2011.

NETTO, Ernesto Junior Silveira. **A Influência da Jurisprudência no Direito Brasileiro. Artigo**.

Disponível em: <<http://www.artigonal.com/>>. Acesso em: 20 abril. 2011.

PENTAHO, PENTAHO CORPORATION, Kettle.
Disponível em: <<http://kettle.pentaho.org/>>. Acesso em: 10 maio. 2011

ROLDÁN, Maria Carina. **Pentaho 3.2 Data Integration: Beginner's Guide**, 2010.

STRECK, Lenio Luiz. **Súmulas no Direito Brasileiro: Eficácia, Poder e Função**. Porto Alegre: Livraria do Advogado, 1998.