Aplicação de um sistema de *Data Warehousing* e de *Business Intelligence* na Biomedicina

Ana Margarida Campos, Benjamin Oliveira, Mário Sequeira, Simão Gonçalves e Susana Marques

Universidade do Minho, Campus de Gualtar, 4710 – 057 Braga, Portugal
a85166@alunos.uminho.pt, PG42815@alunos.uminho.pt,
PG39293@alunos.uminho.pt, PG42850@alunos.uminho.pt
a84167@alunos.uminho.pt

Abstract. Os avanços recentes nas tecnologias da informação facilitam o aumento da capacidade de recolha e armazenamento de dados, sendo os termos *Data Warehouse* e *Business Intelligence* muito mencionados. Nesse sentido, o principal objetivo deste projeto é propor uma arquitetura para Data Warehousing usando um *dataset* previamente disponibilizado com dados de episódios hospitalares e implementar esta arquitetura com a ferramenta *Desktop Power BI* permitindo assim fornecer suporte à decisão clínica, onde os sistemas de informação hospitalar precisam de comunicar para conseguirem compartilhar informações e torná-las disponíveis em qualquer lugar, a qualquer hora.

Keywords: · Data Warehouse · Business Intelligence · Biomedicina · SQL · Desktop Power BI · Decision Support

1 Introdução

A ideia principal de um sistema de *Data Warehouse* consiste em agregar informação proveniente de uma ou mais bases de dados ou de outras fontes para posteriormente a tratar, formatar e consolidar numa única estrutura de dados. É de fundamental importância a implementação de um bom *Data Warehouse*, isto porque, é uma forma de acesso mais facilitada, consistente e legível a toda a informação. O processo de implementação deste passa por, inicialmente, haver fontes de informação heterogéneas, como por exemplo, bases de dados, documentos escritos, entre outros tipos e em seguida, ocorrer uma extração da informação e uma zona de concentração dos mesmos. [1] Através de processos de povoamento de toda a informação anteriormente tratada, tem-se como resultado um *Data Warehouse.*

Posteriormente, poderá ser extraído conhecimento deste para, por exemplo, análises estatísticas, aplicações de apoio à decisão e processos empresariais. Esta extração de conhecimento é muito importante para identificar padrões e suportar estratégias de negócio.

Toda esta informação está estruturada segundo um modelo dimensional (tabela de factos, dimensões e medidas), que poderá ser implementado em estrutura de

estrela (mais espaço e mais eficiência) ou em estrutura de floco de neve (menos espaço e menos eficiência na pesquisa de informação). [1]

É de salientar que, qualquer organização em que o problema é o desempenho na análise dos sistemas de suporte à decisão, como é o caso deste projeto, necessita de um *Data Warehouse*, mas podem ocorrer problemas visto que os sistemas de informação das organizações revelam fraquezas perante a necessidade da análise expedita da informação dos dados segundo várias vertentes e perspectivas de negócio das organizações.

Para tal não acontecer, existem as ferramentas do OLAP (*On-Line Analytic Processing*) que disponibilizam, de um modo rápido e flexível, mecanismos de análise de informação conjugando várias variáveis de negócio. Através destas é possível realizar o tratamento dos dados proveniente de diferentes fontes em tempo real, utilizando métodos mais rápidos e eficazes e permitem também visualizar e organizar dados através dos critérios de seleção pretendidos. No entanto, a maior vantagem do OLAP é a sua capacidade de realizar análises multidimensionais dos dados, associadas a cálculos complexos, análises de tendências e modelação. [1]

Por outro lado, uma das partes mais importantes e poderosas de possuir informação é conseguir transformar a mesma em decisões viáveis, e isso pode ser feito através de *Business Intelligence.*

O objetivo deste é identificar e perceber como é que certos dados têm influência no problema objetivo, e, através dessa interpretação, providenciar melhores serviços de qualidade aos clientes, adotando técnicas apropriadas de gestão. Ou seja, ele permite uma análise de informações consolidadas, a fim de obter soluções específicas ou certas indicações para o processo de decisão. [2]

Concluindo, em conjunto com o uso de um *Data Warehouse*, com o *Business Intelligence* é possível realizar a decisão certa com base nos dados colecionados de diferentes sistemas de informação (neste artigo, referimo-nos aos dados do *dataset* que nos foi disponibilizado), de seguida contextualiza o *dataset* utilizado referindo as principais caraterísticas de cada ficheiro.

Este artigo, após a respetiva introdução atual, está dividido em 5 partes importantes. Começa por uma breve descrição do *Background* aplicável ao *dataset* em uso, explicando como funcionam os sitemas de informação em unidades hospitalares. Depois segue uma contextualização do *dataset* em uso descrevendo cada ficheiro utilizado e a informação retida para responder a tentativas de decisão clínicas. A seguir, é abordado mais profundamente o tópico *Data Warehousing* começando pelo processo *ETL* e focando no seu desenvolvimento neste projeto abordando o modelo dimensional, o tratamento de dados, o povoamento e as estruturas de atualização realizadas para que o *Data Warehouse* fosse criado seguindo todos os critérios fundamentais. Finalmente, através do *Business Intelligence*, foram criadas diversos indicadores que perante a análise ajudam nas decisões clínicas abordadas neste artigo. Concluimos o artigo, mitigando essas mesmas decisões clínicas que se retiraram do projeto.

2 *Background*

Existem diferentes tipos de sistemas de informação nos grandes hospitais de hoje, alguns desses sistemas são complicados (em termos de heterogeneidade e diversidade) e difíceis de gerenciar.

A maioria dos sistemas de saúde é desenvolvida por diferentes pessoas e grupos e construída em diferentes plataformas (arquitetura de sistema, infraestrutura e banco de dados diferentes) o que leva a complicações no que diz respeito à integridade e interoperabilidade entre diferentes sistemas de saúde. A interoperabilidade é a capacidade de diferentes subsistemas acessarem e usarem os dados de forma confiável e rápida de várias fontes, sem a ocorrência de erros, entre os sistemas de saúde e permite que esses sistemas se comuniquem entre si para compartilhar informações apesar destes carecerem de padronização, fazendo a partilha e integração em tais sistemas bastante difícil. [3]

Consequentemente, fornecer serviços de saúde adequados é complicado devido à diversidade, heterogeneidade, dispersão em vários locais e complexidade desses sistemas de saúde. No cenário atual, os pacientes possuem múltiplos registos de saúde em diferentes sistemas de informação, o que significa que as informações relacionadas ao paciente são fragmentadas em diferentes sistemas. [4] A necessidade de ter acesso aos dados do paciente por meio desses sistemas e gerenciar o fluxo de informações entre vários sistemas está aumentando a complexidade dos mesmos e portanto, a interoperabilidade em sistemas de saúde torna-se cada vez mais um requisito do que um recurso.

O sistema de saúde pode ser caracterizado pela presença de um grande número de atores (médicos, enfermeiros, farmacêuticos, administradores, etc.), um grande número de serviços (serviços de atendimento ambulatorial, serviços de reabilitação, etc.) e profissionais altamente especializados.[4] Para melhorar a qualidade do tratamento e garantir diagnósticos mais precisos, as atividades desses atores precisam ser coordenadas e controladas por meio de técnicas de comunicação eficientes. Muitos sistemas de saúde presumem que os utilizadores estão a trabalhar em locais fixos, sem levar em consideração os avanços mais recentes na tecnologia móvel, como telemóveis e dispositivos móveis inteligentes. Assim, o sistema de informação de saúde, em geral, requer uma reengenharia dos seus processos de atendimento.

Por natureza, as informações de saúde são móveis [5], uma vez que as informações do paciente são necessárias para profissionais médicos em diferentes locais para melhorar o diagnóstico, o atendimento eficiente e reduzir os erros médicos. Informações oportunas em situações críticas podem fazer a diferença entre a vida e a morte. Os sistemas de *Data Warehousing* e *Business Intelligence* oferecem uma solução eficaz para os problemas acima mencionados em termos de melhor integração [6], o que, por sua vez, resulta num acesso mais eficiente aos diferentes sistemas de saúde e permite a aquisição de dados médicos do paciente e o controle de informações de fontes múltiplas, melhorando assim o atendimento ao paciente.

3 Contextualização

Neste projeto usou-se um *dataset* que contém diferentes conjuntos de dados separados por diferentes ficheiros *.csv* focando-se em episódios de urgências clínicas:

- **urgency_exams:** Este ficheiro contém informações relativas sobre cada exame que o paciente realizou num determinado episódio, apresentando uma descrição e o número do exame, podendo verificar-se que existem exames diferentes com a mesma descrição, algo que deve ser considerado relevante, numa seguinte fase de modelação.
- **urgency_prescriptions:** Este ficheiro aborda informações sobre uma determinada prescrição para um determinado episódio: o seu código, o profissional que a prescreve, a data que é realizada e uma breve descrição sobre a mesma. E também menciona qual o código do medicamento prescrito, a sua quantidade e uma breve descrição sobre o medicamento em si.
- **urgency_procedures:** Neste ficheiro estão inseridas informações relativas a intervenções, tais como qual a data de prescrição da mesma, quem a prescreveu, qual é a data de começo, breves indicações sobre a intervenção e se foi cancelada quem a cancelou e quando.
- **urgency_episodes_new:** Neste ficheiro encontram-se informações relativas sobre o paciente de um determinado episódio de urgência: a sua data de nascimento, o sexo, o distrito, a data da sua admissão, o profissional que o admitiu, as causas externas para a sua admissão, o profissional que realizou a triagem e a respetiva data, a escala de dor que lhe foi atribuída, assim como a cor, o diagnóstico que lhe foi feito, a data de quando esse diagnóstico foi realizado e qual o profissional que o realizou, o destino que o paciente teve após o episódio de urgência e o profissional que lhe deu alta, em conjunto com a respetiva data e razão.
- **icd9_hierarchy:** Este ficheiro possuí uma lista de códigos que servem para a classificação de doenças, onde o código de nível 1 classifica a doença de uma forma mais geral, e ao progredirmos na escala de níveis, encontraremos mais especificações sobre tal doença. A informação encontra-se de forma hierárquica e relaciona-se com o código de diagnóstico do ficheiro *urgency_episodes_new*.

Através deste *dataset*, pretende-se conseguir encontrar indicadores que permitam ajudar na decisão de, por exemplo, quantos profissionais de saúde deverão encontrar-se a trabalhar nas urgências em determinada altura do ano, quais são as doenças que afetam mais as urgências nas admissões dos pacientes, qual é a gravidade dessas doenças, entre outras.

4 Data Warehousing

4.1 ETL

Para a construção de um *Data Warehousing* são necessários diferentes passos principalmente ao nível da extracção e processamento de dados. O processo ETL destina-se à extracção e transformação dos dados e termina com a inclusão destes no *Data Warehousing*. Esta fase caracteriza-se por englobar procedimentos de limpeza, integração e transformação de dados. Segundo a literatura este é o processo mais crítico e demorado na construção de um *Data Warehouse*.

A figura seguinte descreve de forma geral o processo de ETL que se divide em três fases cruciais: Extração; Transformação; Carga. [7]

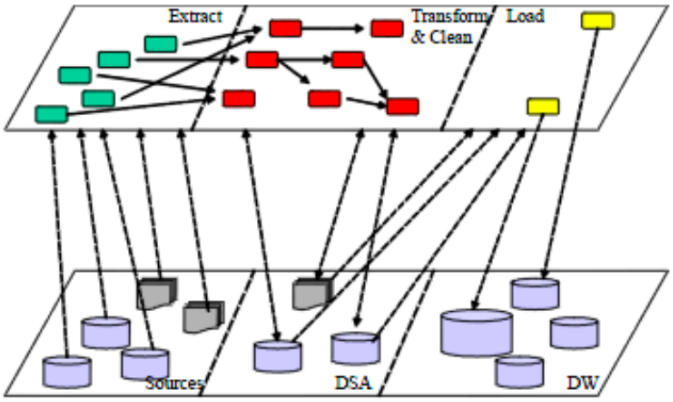

Fig. 1. Ilustração do processo de ETL [8]

Fase de Extração: A camada inferior representa o armazenamento dos dados que são utilizados em todo o processo. No lado esquerdo pode-se observar os dados "originais" provenientes, na maioria dos casos, de bases de dados ou, então, de ficheiros heterogéneos, por exemplo de texto. Os dados provenientes destas fontes são obtidos (como é ilustrado na área superior esquerda da figura 2), por rotinas de extracção que fornecem informação igual ou modificada, relativamente à fonte de dados original.[9]

Fase de Transformação: Posteriormente, esses dados são propagados para a *Data Staging Area* (DSA) onde são transformados e limpos antes de serem carregados para o *Data Warehouse*. O nível de alteração e manipulação dos dados extraídos centra-se na qualidade dos mesmos, assim como nas necessidades de utilização. Desta forma, é efetuada uma limpeza dos dados irrelevantes, de informações duplicadas, remoção de erros e correção de dados perdidos. [9] Neste projeto, a estratégia utilizada nesta fase assentou no que é descrito na secção relativa ao Tratamento de Dados.

Fase de Carga: O carregamento dos dados, é realizado através das actividades de carga representadas na parte superior direita da figura. Esta fase, direcionada para o contexto do desenvolvimento do *Data Warehouse* deste artigo em concreto, é explorada na secção relativa ao Povoamento.

4.2 Modelo Dimensional

Numa primeira fase, procedeu-se à análise do *dataset* de modo a arquiteturar a relevância de certas informações em prol de outras.

Percebeu-se que a melhor estratégia a adotar seria a elaboração de um modelo dimensional em floco de neve, visto que a constituição do mesmo requer uma tabela de factos e tabelas de dimensão hierarquizadas.

Assim, considerou-se a tabela *facts_episodio* como sendo a tabela de factos, uma vez que neste *dataset* se pretende tratar da existência de episódios de urgência e todos os atributos que se relacionam a estes episódios.

Após uma análise detalhada dos ficheiros *.csv* do *dataset* escolhido, foi desenvolvido o seguinte modelo dimensional:

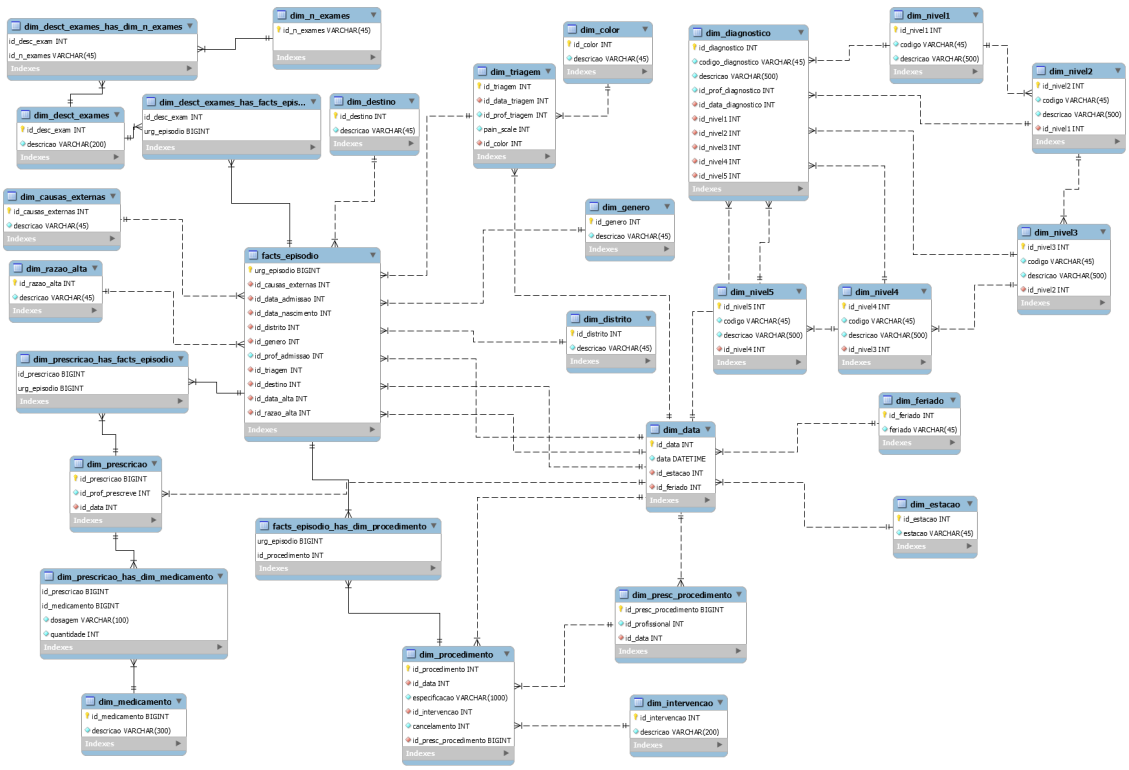

Fig. 2. Modelo Lógico

De modo a analisar em pormenor todas as tabelas do modelo, estas são específicadas de seguida:

- ***facts_episodio:*** Tabela de factos que representa os episódios de urgência. Tem associada a si o número de identificação do episódio de urgência, o número de identificação do professional que deu admissão ao episódio de urgência e as chaves estrangeiras das tabelas de dimensão: *dim_causas_externas*, *dim_data*, *dim_distrito*, *dim_genero*, *dim_triagem*, *dim_destino* e *dim_razao_alta*.
- ***dim_prescricao:*** Tabela de dimensão que apresenta a data de quando foi feita a prescrição, qual foi o profissinal que a fez e o número de identificação da prescrição. Tem associada a si a chave estrangeira da tabela *dim_data*.
- ***dim_prescricao_has_facts_episodio:*** Tabela de dimensão que relaciona o episódio de urgência com a prescrição.
- ***dim_medicamento:*** Tabela de dimensão que apresenta o número de identificação do medicamento e a sua descrição.
- ***dim_prescricao_has_dim_medicamento:*** Tabela de dimensão que relaciona uma prescrição com os medicamentos e que apresenta a dosagem e quantidade dos medicamentos.
- ***dim_procedimento:*** Tabela de dimensão que representa um procedimento. Tem associada a si o número de identificação, a especificação, o cancelamento do procedimento, e as chaves estrangeiras das tabelas: *dim_data*, *dim_intervencao* e *dim_presc_procedimento*.
- ***facts_episodio_has_dim_procedimento:*** Tabela de dimensão que relaciona um episódio de urgência com um procedimento.
- ***dim_intervencao:*** Tabela de dimensão que apresenta o número de identificação de uma intervenção e a sua descrição.
- ***dim_presc_procedimento:*** Tabela de dimensão que representa a prescrição dos procedimentos. Tem associada a si o número de identificação do professional que a fez e a chave estrangeira da tabela *dim_data*.
- ***dim_data:*** Tabela de dimensão que representa a data. Tem associada a si as chaves estrangeiras das tabelas: *dim_feriado* e *dim_estacao*.
- ***dim_estacao:*** Tabela de dimensão que representa a estação do ano.
- ***dim_feriado:*** Tabela de dimensão que representa os feriados do ano.
- ***dim_distrito:*** Tabela de dimensão que representa o distrito do utente.
- ***dim_genero:*** Tabela de dimensão que representa o género do utente.
- ***dim_diagnostico:*** Tabela de dimensão que representa o diagnóstico resultante do episódio de urgência. Tem associada a si o seu número de identificação, o código e a descrição do diagnóstico, o número de identificação do professional que deu o diagnóstico e as chaves estrangeiras das tabelas: *dim_data*, *dim_nivel1*, *dim_nivel2*, *dim_nivel3*, *dim_nivel4* e *dim_nivel5*.
- ***dim_nivel1:*** Tabela de dimensão que representa a descrição de nível 1 do diagnóstico (a menos específica).
- ***dim_nivel2:*** Tabela de dimensão que representa a descrição de nível 2 do diagnóstico (mais específica que o nível 1 e menos que o nível 3).
- ***dim_nivel3:*** Tabela de dimensão que representa a descrição de nível 3 do diagnóstico (mais específica que o nível 2 e menos que o nível 4).

- ***dim_nivel4:*** Tabela de dimensão que representa a descrição de nível 4 do diagnóstico (mais específica que o nível 3 e menos que o nível 5).
- ***dim_nivel5:*** Tabela de dimensão que representa a descrição de nível 5 do diagnóstico (a mais específica).
- ***dim_razao_alta:*** Tabela de dimensão que representa a razão da alta.
- ***dim_causas_externas:*** Tabela de dimensão que representa causas externas.
- ***dim_destino :*** Tabela de dimensão que representa o destino do utente.
- ***dim_desct_exames:*** Tabela de dimensão que representa o número de identificação do exame e a sua descrição.
- ***dim_desct_exames_has_facts_episodio:*** Tabela de dimensão que relaciona o número de exames realizados com a descrição de cada um.
- ***dim_n_exames:*** Tabela de dimensão que representa o número de exames realizados.
- ***dim_desct_exames_has_dim_n_exames:*** Tabela de dimensão que relaciona os exames com o episódio de urgência.
- ***dim_triagem:*** Tabela de dimensão que representa a triagem realizada ao utente. Tem associada a si o número de identificação da triagem e do professional que a realizou, a escala de dor e as chaves estrangeiras das tabelas: *dim_data* e *dim_color*.
- ***dim_color:*** Tabela de dimensão que representa a cor que foi atribuída ao utente na triagem.

Neste esquema são ainda visíveis tabelas de relações de N para N que possuem sempre duas chaves estrangeiras primárias que apontam para as chaves primárias das duas tabelas principais.

4.3 Tratamento de dados

O tratamento e dados considera-se uma das fases mais importantes, uma vez que se destaca pelo empenho e dedicação que carece, desde a procura por erros, até à descoberta da melhor estratégia para a resolução destes.

Como o *dataset* é composto por múltiplos ficheiros com uma elevada quantidade de dados, o cuidado no tratamento destes dados terminou por ser um processo desafiante e árduo.

Primeiramente, removeu-se do *dataset* colunas com informação considerada irrelevante face à análise deliberada aquando do início da elaboração do projeto por apresentarem pouca informação importante para o mesmo. As colunas eliminadas foram: DIAGNOSIS_NOTES do ficheiro *urgency_episodes_new*, as colunas PVP e COMPARTICIPATION do ficheiro *urgency_exams* e as colunas DT_CANCEL e NOTE_CANCEL do ficheiro *urgency_procedures*. Estas duas últimas colunas foram eliminadas após uma decisão ponderada sobre a data ou a razão de um cancelamento numa intervenção, no projeto, não ser informação importante para o mesmo, considerando que é apenas importante saber se um cancelamento ocorreu, para uma maior qualidade na contagem de intervenções que realmente foram efetuadas (daí não se ter apagado a coluna

ID_PROFISSIONAL_CANCEL, mas substituído todos os seus valores em campos não preenchidos (nulos) por 0, para mais à frente se poder usar a mesma como um procedimento que permitirá saber se uma intervenção foi cancelada).

A seguir o foco centrou-se na procura de todos os valores nulos, valores com informação incompleta e, até mesmo, valores corrompidos (isto é, com caracteres que não faziam sentido no local onde se encontravam).

Para o tratamento de valores nulos e de informação incompleta, colocaram-se todos esses valores a "NA". Ainda dentro de valores corrompidos, subtstituiu-se o ponto e vírgula por apenas vírgula.

Relativamente a datas e horas, todas elas foram convertidas no formato: aaaa-mm-dd hh-mm-ss; e para lidar com o tratamento de valores nulos em colunas deste tipo, sugeriu-se adotar o seguinte padrão: 9999-01-01 00:00:00, uma vez que não faria sentido alguma data correta encontrar-se no ano 9999.

4.4 Povoamento

O passo seguinte neste artigo consistiu em povoar o modelo desenvolvido. Para tal, e seguindo o trabalho desenvolvido até ao ponto atual, utilizamos os dados processados no ponto 4.3.

Para a inserção da informação na base de dados usaram-se duas metodologias diferentes. Numa primeira fase foi explorada a inserção dos dados através de *SQL*. Começamos por efetuar um carregamento dos ficheiros disponibilizados para a base de dados (um total de cerca de 280 mil linhas, obtidas a partir de 5 ficheiros diferentes) e, só depois de termos estes dados em bruto na base de dados, procedemos á distribuição da informação (inserir os dados corretos nas tabelas corretas). Para tal recorremos a *querys* sobre o *dataset*, inserindo os dados obtidos nas tabelas correspondentes.

```
-- insercao na dimensão dim_triagem
insert into dim_triagem(id_data_triagem, id_prof_triagem, pain_scale, id_color)
select distinct a1.id_data, a2.ID_PROF_TRIAGE, a2.PAIN_SCALE, a3.id_color
    from urgency_episodes_new a2
    left join dim_data a1 on a1.data = a2.DT_ADMITION_TRAIGE
    left join dim_color a3 on a3.id_color = a2.ID_COLOR;
```

Fig. 3. Exemplo de povoação usando SQL.

Foi também explorada a inserção de informação na base de dados utilizando Java. Para tal foi desenvolvido um script responsável por carregar os ficheiros em *"runtime"*, dando *parse* dos mesmos e inserindo a informação nas tabelas correspondente.

```
if (cells[2].contains("'") && cells[1].contains("'"))
    s = "INSERT IGNORE INTO dim_desct_exames_has_dim_n_exames (id_desc_exam, id_n_exames) VALUES" +
            "((select id_desc_exam from dim_desct_exames where descricao = \""+ cells[2]+"\")," +
            "(select id_n_exames from dim_n_exames where id_n_exames = \""+ cells[1]+"\"))";
```

Fig. 4. Exemplo de povoação usando Java.

Após este processo estar finalizado, todas as 28 tabelas se encontravam (corretamente) populadas e prontas a serem utilizadas nas etapas seguintes.

4.5 Estruturas de atualização

Com a finalidade de tornar o projeto mais enriquecedor e futuramente modificável, foram implementadas algumas estruturas de atualização (*triggers*, *procedures* e *functions*) consideradas necessárias para o desenvolvimento do mesmo. Estas são apresentadas de seguida.

– Foi criada uma *function* que tem como objetivo associar as diferentes datas às estações do ano, ou seja, Inverno, Primavera, Outono e Verão.

```
DELIMITER //
DROP FUNCTION IF EXISTS Estacao //
CREATE FUNCTION `Estacao` ( data datetime )
RETURNS varchar(100)
READS SQL DATA
DETERMINISTIC
BEGIN

    DECLARE estacao varchar(100);

    IF data = '9999-01-01 00:00:00' THEN
     SET estacao = 'NA';
    ELSEIF DATE_FORMAT(data, '%m-%d') >= '03-20' and
     DATE_FORMAT(data, '%m-%d') < '06-21' THEN
        SET estacao = 'Primavera';

    ELSEIF DATE_FORMAT(data, '%m-%d') >= '06-21' and
     DATE_FORMAT(data, '%m-%d') < '09-22'  THEN
        SET estacao = 'Verao';

    ELSEIF DATE_FORMAT(data, '%m-%d') >= '09-22' and
     DATE_FORMAT(data, '%m-%d') < '12-21' THEN
        SET estacao = 'Outono';

    ELSE
        SET estacao = 'Inverno';

    END IF;

    RETURN estacao;
```

```
END; //
```

– Como consequência da *function* anterior, foi necessária a criação de um *trigger* que possui como finalidade acionar a *function* antes da inserção de novos registos relativos à data no *Data Warehouse*. Permite relacionar a cada data da dimensão *dim_data*, o índice associado à estação correspondente.

```
DELIMITER //
DROP TRIGGER IF EXISTS estacoes //
CREATE TRIGGER estacoes
BEFORE INSERT
ON dim_data FOR EACH ROW
BEGIN
   set new.id_estacao = (select id_estacao from dim_estacao
    where estacao = Estacao(new.data));
END;//
DELIMITER ;
```

– De maneira muito similar à *function* que relaciona estações do ano às datas, foi implementada outra, mas com o intuito de verificar se uma determinada data é um feriado português. Os feriados considerados no projeto foram apenas feriados fixos, ou seja, que não mudam de ano para ano. Estes foram os seguintes: ano novo, dia da liberdade, dia do trabalhador, dia de Portugal, implementação da república, dia de todos os santos, restauração da independência e, por último, o Natal.

```
DELIMITER //
DROP FUNCTION IF EXISTS Feriado //
CREATE FUNCTION `Feriado` ( data datetime )
RETURNS varchar(100)
READS SQL DATA
DETERMINISTIC
BEGIN
  DECLARE feriado varchar(100);
   IF data = '9999-01-01 00:00:00' THEN
      SET feriado = 'NA';
   ELSEIF DATE_FORMAT(data, '%m-%d') ='01-01' THEN
      SET feriado = 'Dia de Ano-Novo';
  ELSEIF DATE_FORMAT(data, '%m-%d') = '04-25' THEN
      SET feriado = 'Dia da Liberdade';
   ELSEIF DATE_FORMAT(data, '%m-%d') = '05-01' THEN
      SET feriado = 'Dia do Trabalhador';
   ELSEIF DATE_FORMAT(data, '%m-%d') = '06-10' THEN
      SET feriado = 'Dia de Portugal';
   ELSEIF DATE_FORMAT(data, '%m-%d') = '10-05' THEN
      SET feriado = 'Dia da Implementa  o da Rep blica';
   ELSEIF DATE_FORMAT(data, '%m-%d') = '11-01' THEN
      SET feriado = 'Dia de Todos os Santos';
   ELSEIF DATE_FORMAT(data, '%m-%d') = '12-01' THEN
```

```
        SET feriado = 'Restaurac o da Independ ncia';
    ELSEIF DATE_FORMAT(data, '%m-%d') = '12-25' THEN
        SET feriado = 'NATAL';
   ELSE
        SET feriado = 'NA';
   END IF;
    RETURN feriado;
END; //
DELIMITER ;
```

– De modo a usufruir da função enunciada em cima, foi necessária a criação de um *trigger* que, antes da inserção de datas na dimensão *dim_data*, é acionado e coloca o identificador de feriado em cada uma.

```
DELIMITER //
DROP TRIGGER IF EXISTS feriados //
CREATE TRIGGER feriados
BEFORE INSERT
ON dim_data FOR EACH ROW
BEGIN
   set new.id_feriado = (select id_feriado from dim_feriado
     where feriado = Feriado(new.data));
END;//
DELIMITER ;
```

– Com vista a identificar possíveis cenários de cancelamento dos procedimentos, procedeu-se à implementação de um *procedure* que coloca o atributo *cancelamento* a 1, caso o mesmo tenha sido efetuado. Este utiliza como auxílio o *ID_PROFESSIONAL_CANCEL* do ficheiro *.csv* relativo aos procedimentos, que indica o profissional que procedeu à suspensão. Caso o identificador relativo ao profissional seja diferente de zero, então é convertido para o valor 1 indicando que houve o cancelamento.

```
DELIMITER //
CREATE PROCEDURE cancelado()
BEGIN
   UPDATE dim_procedimento
   SET cancelamento = 1
   WHERE dim_procedimento.cancelamento != 0;
END //
DELIMITER ;
```

– A última estrutura de atualização criada permite auxiliar a povoação da dimensão *dim_diagnostico*. O código de um diagnóstico possui cinco níveis distintos sendo o primeiro nível o menos específico e, à medida que o nível aumenta, a especificação aumenta consequentemente. De modo a interligar

todos estes níveis com o diagnóstico em causa, foi desenvolvido um *trigger* que é acionado antes da inserção dos dados. Este começa por verificar se o código em causa se insere no nível cinco (nível mais específico). Caso isto aconteça, então é inserido o índice relativo ao nível e, a povoação dos restantes níveis é possível visto que estes se relacionam entre si. O código repete-se da mesma maneira para todos os outros níveis. De seguida é apenas apresentado uma parte do *trigger* uma vez que este é bastante complexo e, por isso, muito grande.

```
    DELIMITER //
DROP TRIGGER IF EXISTS povoarDiagnostico //
CREATE TRIGGER povoarDiagnostico
BEFORE INSERT
ON dim_diagnostico FOR EACH ROW
BEGIN
  if  ( exists (select level_5_code from icd9_hierarchy
   where  new.codigo_diagnostico = level_5_code )) = 1
   then
        set new.id_nivel5 = (select id_nivel5 from
   dim_nivel5 where codigo =
     (select level_5_code from icd9_hierarchy where new.
   codigo_diagnostico = level_5_code limit 1));
        set new.id_nivel4 = (select id_nivel4 from
   dim_nivel5 where codigo =
     (select level_5_code from icd9_hierarchy where new.
   codigo_diagnostico = level_5_code limit 1));
        set new.id_nivel3 = (select id_nivel3 from
   dim_nivel4 where id_nivel4 = new.id_nivel4);
        set new.id_nivel2 = (select id_nivel2 from
   dim_nivel3 where id_nivel3 = new.id_nivel3);
        set new.id_nivel1 = (select id_nivel1 from
   dim_nivel2 where id_nivel2 = new.id_nivel2);
```

5 Business Intelligence

Nesta etapa, após o desenvolvimento do *Data Warehouse* e usando a ferramenta *Power BI Desktop*, apresenta-se os indicadores principais e uma respetiva análise dos mesmos que permita um suporte à decisão clínica.

5.1 Indicadores e Análise

Neste primeiro indicador analisamos a quantidade de pacientes que ocorrem nas urgências numa determinada estação do ano.

Inferimos que existe uma maior quantidade de admissões durante o Verão e a Primavera do que no Inverno ou Outono. Estes resultados foram surpreendentes, pois seria mais expectável que a estação mais afetada seria o Inverno ou o Outono devido a uma maior facilidade de contágio de vírus. Inferiu-se que uma das razões para esta contagem no Verão poderá ser um aumento fluxo de pessoas na rua, havendo maior percentagem de terem acidentes.

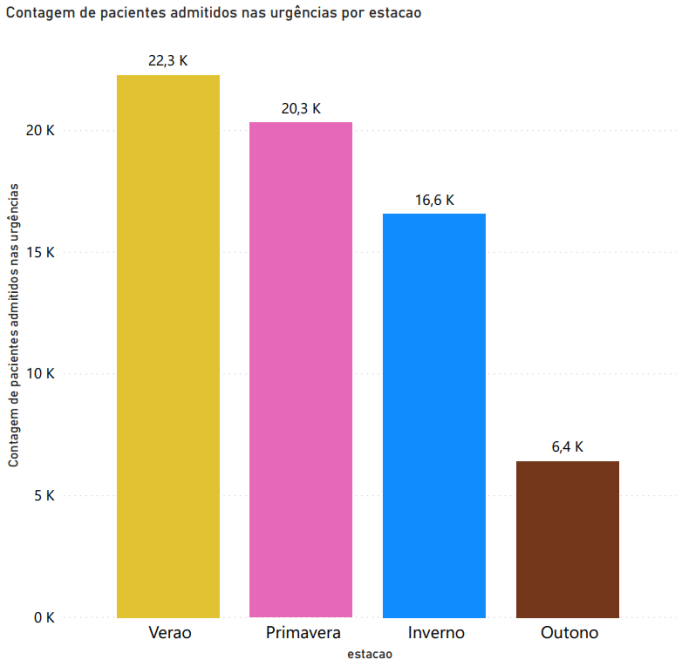

Fig. 5. Admissões por estação.

Na figura seguinte analisamos as causas mais comuns de internamento de nível 1 (mais abrangente), em relação à estação do ano em que foram registadas.

Observamos que a causa que se destaca mais e consistentemente pelas quatro estações é *Symptoms, signs and ill-defined conditions*, sendo espectável ,já que é o que representa as causas relacionadas com todo o tipo de doenças.

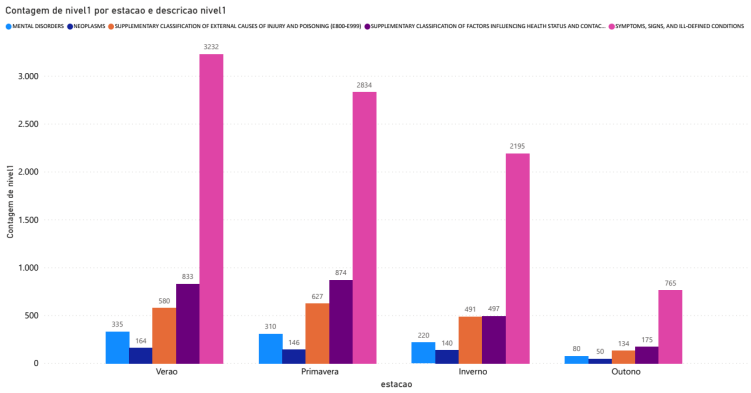

Fig. 6. Descrições de nível 1 por estação.

Neste gráfico, analisamos as causas de nível 5 (mais especificas) tendo em conta as ilações tiradas no gráfico acima representado. Assim comparamos as causas de nível 5 com as de nível 1.

Vemos que as doenças ou sintomas mais comuns nos casos de urgência são Dores no Peito, seguido de casos em que os pacientes apresentam Diarreia.

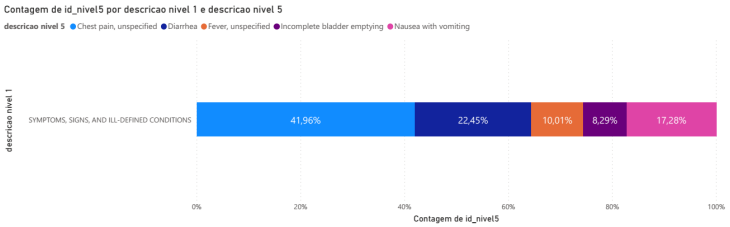

Fig. 7. .

No gráfico abaixo representado, podemos observar uma comparação entre o género Feminino e Masculino na contagem de episódios e no destino de cada paciente.

Em termos do número de episódios, o género Feminino conta com mais casos que o Masculino, mas podemos ver que grande maioria é encaminhada para a Unidade Maternidade Júlio Dinis, ao passo que o género Masculino é encaminhado para o serviço de Urologia.

De notar que a Unidade Maternidade Júlio Dinis é exclusiva ao género Feminino no gráfico, como esperado, podendo-se inferir que várias urgências a nível feminino ocorrem quando as mulheres dão entrada no hospital para realizarem um parto.

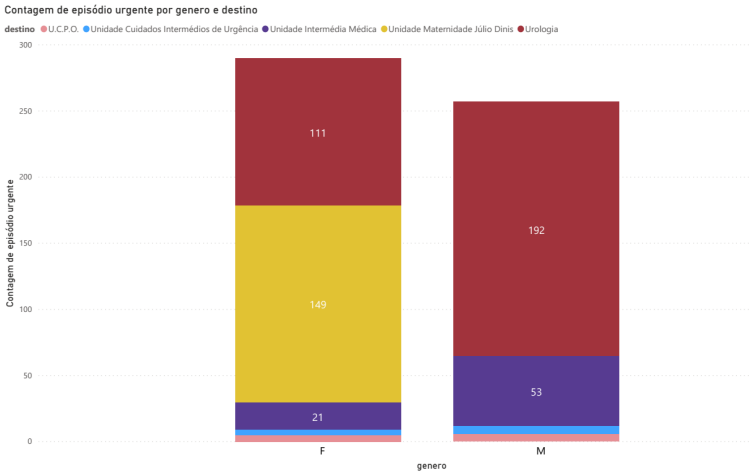

Fig. 8. Admissões por género e destino.

O seguinte indicador foi desenvolvido para analisar a correlação entre a cor atribuída a cada paciente e a escala de dor sentida pelo mesmo.

Analisando o indicador, e ordenando as cores por Verde, Amarelo e Laranja, conseguirmos ver que os níveis de dor mais baixos (1,2,3) estão associados à cor Verde, os níveis de intermédios (4,5,6) à cor Amarelo e os níveis mais altos (7,8,9,10) à cor Laranja.

Assim concluímos que aos pacientes lhes é atribuído uma cor de acordo com o nível de dor que sentem.

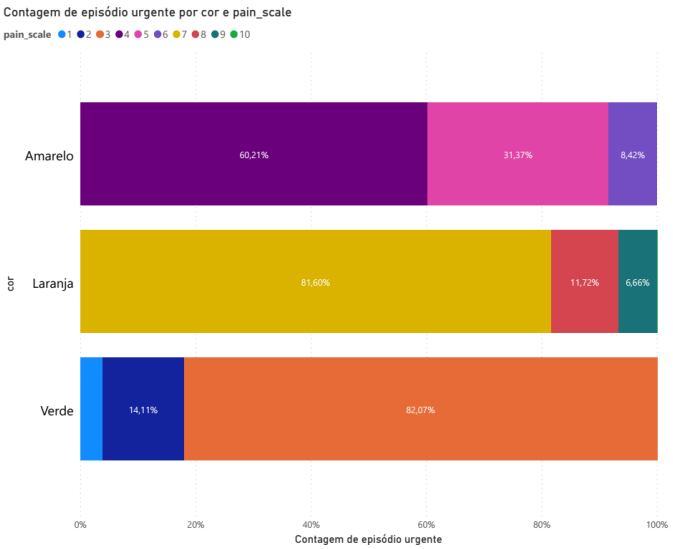

Fig. 9. Analise da escala de dor por cor.

No presente indicador observamos uma tabela de causas mais frequentes nos episódios de urgência.

Conseguimos retirar dos gráficos que a causa mais comum de episódios registados é a de Doença, seguido de Queda.

Causas externas	Contagem de episódio urgente
Doenca	58405
Queda	3429
Agressao	686
Acidente Pessoal	262
Intoxicacao	114
Outras	77
Acidente Escolar	57
Queimadura	26
Gravidas E Parturientes	3
Queimadura Solar	1
Total	**63060**

Fig. 10. Principais causas de admissão.

O indicador seguinte permite-nos observar quais as descrições mais comuns para cada exame efetuado.

À partida observamos que a descrição com maior número de ocorrências em exames é a Transição dorso-lombar, duas incidências.

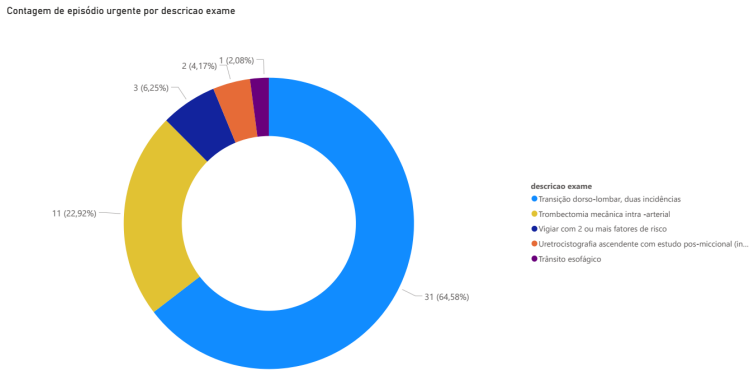

Fig. 11. Descrições de exames efetuados.

O gráfico que se segue representa as razões dadas aquando da alta do paciente.

Constatamos que a maioria dos pacientes tem alta para o domicilio, sendo que só uma pequena percentagem é encaminhada para um Centro de Saúde ou para uma consulta externa.

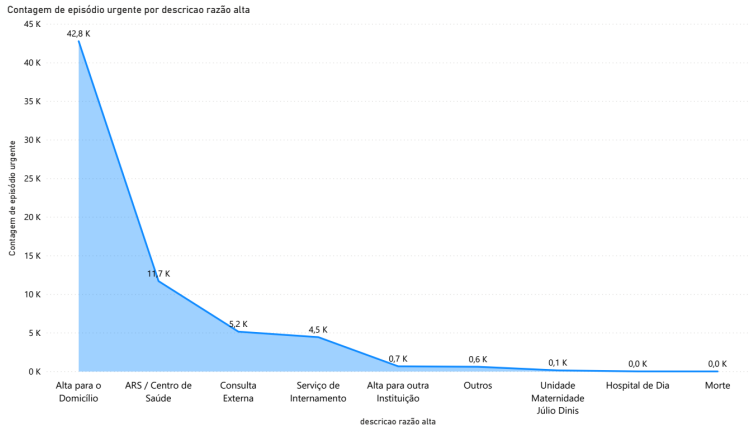

Fig. 12. Razões para alta de um paciente.

Neste gráfico observamos o número de pessoas que foram admitidas nas urgências nos 5 feriados descritos na legenda: Dia da Liberdade, Dia da Implementação da Republica, Dia do Trabalhador, Dia de Portugal e Dia de Ano-Novo.

Após a analise do gráfico concluímos que apesar de os números de admissão serem mais altos no Dia da Liberdade e no Dia da Implementação da Republica, a diferença não é significativa para retirarmos conclusões relevantes em relação à afluência de pessoas às urgências.

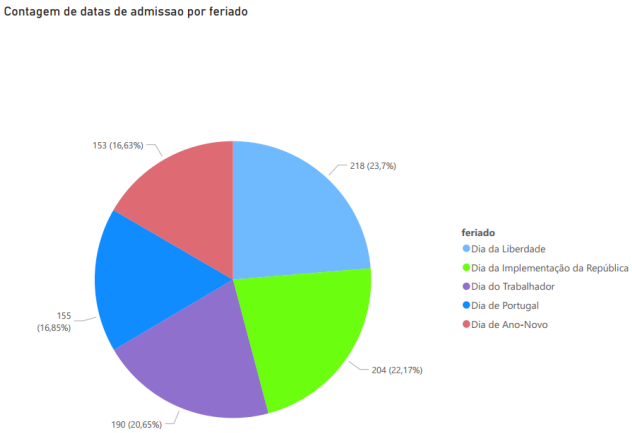

Fig. 13. Análise do número de admissões de pacientes em feriados

Pela analise deste gráfico, conseguimos inferir que os medicamentos mais prescritos nas urgências são os *Zolpidem* de 10 *mg*, sendo a caixa mais utilizada a de 14 unidades, seguida pela de 20 unidades.

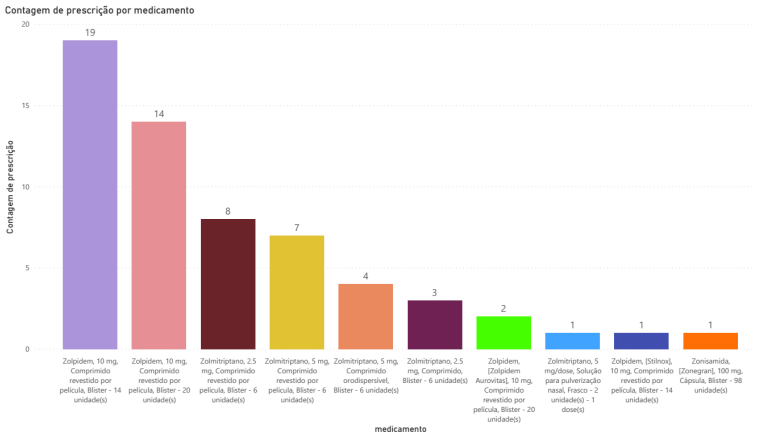

Fig. 14. Análise dos medicamentos prescritos.

6 Conclusão

Este artigo teve como objetivo fornecer as informações básicas e outras mais importantes para o domínio médico usando tecnologias apropriadas tais como o *SQL* e o *Power BI Desktop*.

Analisando os resultados obtidos conseguiu-se concluir que durante o verão ocorreu uma maior admissão nas urgências sendo o tipo de diagnóstico mais comum *Sympyoms,signs and ill-defined conditions*. O diagnóstico mais específico e de maior abrangência relacionado com o mesmo é *Chest pain, unspecified*. É possível também analisar que ocorreu uma maior incidência do género feminino nas urgências devido, particularmente, a questões de maternidade. Com os dados fornecidos foi também permitido observar que a principal causa externa é a de doença e em seguida quedas e relacionar as diferentes cores da triagem (nomeadamente verde, amarelo e laranja) com diferentes níveis de dor. De acordo com informações relativas a exames médicos, é possível inferir que o exame a qual mais se recorreu diz respeito a transição dorso-lombar, duas incidências. A principal razão de alta relaciona-se a alta para o domicilio e o medicamento mais requerido e prescrito é relativo ao *Zolpidem*, 10 *mg* (comprimido revestido por película, *Blister* - 14 unidade(s)). Em relação aos feriados considerados, nota-se que existiu uma afluência nas urgências semelhante em todos.

Em suma, com a elaboração deste projeto conseguiu-se extrair informações fundamentais para o apoio à decisão médica, de acordo com o objetivo pretendido.

Referências

1. Martinho, B., Santos, Maribel.: An Architeture for Data Warehousing in Big Data Environments - ALGORITMI Research Centre, University of Minho
2. S. Alcântara. Business Intelligence (BI) como Auxiliar à Gestão do Negócio. Master's thesis, Faculdade de Tecnologias da Zona Leste, São Paulo, 2010
3. Cardoso, L., Marins, F., Portela, F., Abelha, A., Machado, J.: Healthcare Interoperability through Intelligent Agent Technology, CENTERIS 2014 - Conference on ENTERprise Information Systems / ProjMAN, 2014
4. Al-Sakran, H.: Framework Architecture for Improving Healthcare Information Systems Using Agent Technology. International Journal of Managing Information Technology 7(1), 2015
5. J. Machado, V. Alves, A. Abelha, and J. Neves. Ambient intelligence via multiagent systems in medical arena. International Journal of Engineering Intelligent Systems, Special issue on Decision Support Systems, 2007.
6. A. Cabral, A. Abelha, M. Salazar, C. Quintas, F. Portela, J. Machado, J. Neves, and M. Santos. Knowledge Acquisition Process for Intelligent Decision Support in Critical Health Care. IGI Global Book, 2013.
7. Kimbal, J.: The Data Warehouse ETL Toolkit.- John Wiley Sons, 2004.
8. Panos Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., Skiadopoulos, S.: A generic and customizable frameworkfor the design of ETL scenarios. Information Systems, 2005
9. R. Kimbal, L. Reeves, M. Ross, W. Thornthwaite. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying DataWarehouses. John Wiley Sons, February 1998.