

Mestrado em Engenharia Informática e de Computadores

SIAD - Sistemas de Informação para Apoio à Decisão

Assunto:	Trabalho de SIAD – BIT (Bikes In Transit) - Fase 1 e 2
Departamento:	DEETC
Última Revisão:	2013/05/05

Resumo:

Controlo de versões:

Grupo nº	7066	Frederico Ferreira		
	14940	Angelo Borges		
	32342	Rui Miranda		
Versão	Autor	Revisão	Data	Observações
1.0	Grupo	Grupo	2013/03/09	
1.1	Grupo	Grupo	2013/04/05	
1.2	Grupo	Grupo	2013/05/06	

Índice

1. SÍNTESE DO CASO	5
1.1. PRIMEIRA FASE	5
1.1.1. Requisitos Mínimos.....	5
1.1.2. Objectivos.....	5
1.2. SEGUNDA FASE.....	6
1.2.1. Requisitos Mínimos.....	6
1.2.2. Objectivos.....	6
1.3. NOTAS PARA A IMPLEMENTAÇÃO	7
2. ANÁLISE DO PROBLEMA	8
2.1. MODELO LÓGICO DO SISTEMA OPERACIONAL.....	8
2.2. CLASSIFICAÇÃO DAS ENTIDADES	8
3. ARQUITECTURA DA SOLUÇÃO.....	9
3.1. PROCESSO DE NEGÓCIO “OPERACIONAL”	10
3.1.1. Identificação de Hierarquias	10
3.1.2. Produção dos modelos multidimensionais.....	10
3.1.3. Data Data Profiling e Assunções	11
3.1.4. Arquitetura do Datamart “Operacional”	13
3.1.5. TABLE Operacoes.Morada	14
3.1.6. TABLE Operacoes.TipoServico	14
3.1.7. TABLE Operacoes.Cliente	14
3.1.8. TABLE Operacoes.Funcionario	15
3.1.9. TABLE Operacoes.Data.....	17
3.1.10. TABLE Operacoes.Hora	17
3.1.11. TABLE MonthDescr	17
3.1.12. TABLE DiaSemanaDescr	18
3.1.13. TABLE Operacoes.Servicos	18

3.1.14.	TABLE Operacoes.Entregas	19
3.1.15.	Modelo do DATAMART Operacional	20
3.1.	PROCESSO DE NEGÓCIO "FINANCEIRO"	21
3.1.1.	TABLE Operacoes.Meses	22
3.1.2.	TABLE Operacoes.Rubrica	22
3.1.3.	TABLE Operacoes.Financeira	23
3.1.4.	Modelo do DATAMART Financeiro	23
4.	DESENVOLVIMENTO E IMPLEMENTAÇÃO DO ETL.....	24
4.1.	DESCRIÇÃO DA INTEGRAÇÃO	24
4.2.	ANEXOS	25
4.2.1.	Configuração	25
5.	CONSIDERAÇÕES FINAIS:.....	27
5.1.	GERAL.....	27
5.2.	MELHORIAS	27
5.3.	FUTURO	28

1. Síntese do Caso

A BIT (Bikes In Transit) é uma empresa de entregas portuguesa que opera em Lisboa e no Porto. Nasceu em 2009 e tem como característica principal e diferenciadora, a utilização exclusiva da bicicleta como veículo de entregas.

Da sua carteira de clientes fazem parte um conjunto de empresas e clientes particulares. Especializou-se em entregas de documentos e de outros pequenos volumes.

Em 2010 iniciou a implementação de um sistema informático e actualmente, este assenta, essencialmente sobre um sistema OLTP. Algumas actividades, como a gestão dos custos fixos, são desenvolvidas em folhas de cálculo. Para auxiliar as tomadas de decisão, táticas e estratégicas, a empresa decidiu investir na melhoria desse sistema, através da implementação de um data warehouse, que integra toda a informação.

1.1. Primeira Fase

1.1.1. Requisitos Mínimos

O modelo multidimensional a desenvolver deve dar suporte a questões relacionadas com as entregas de encomendas e com a gestão da empresa. As análises pretendidas consistem na observação da evolução das entregas por tipo, loja, região, freguesia, cliente¹, ano, mês, trimestre, semana, dia da semana e hora do dia. Serão efectuadas análises sobre o número de entregas por estafeta, quantas ficam pendentes e o qual número médio de tentativas necessárias para efectuar a entrega. É também importante saber o número de entregas programadas cujo horário não é cumprido.

Pretende-se também disponibilizar informação sobre o balanço mensal, que engloba os custos salariais (em bruto), os custos fixos² e o valor dos pagamentos, pretendendo-se efectuar análises por mês, trimestre, ano, tipo de custo.

1.1.2. Objectivos

1. Identificar os processos de negócio mais relevantes para os requisitos pretendidos;

¹ Normalmente os clientes são segmentados segundo os seus rendimentos. Geralmente assume-se a categorização nas classes alta (A), média-alta (B), média (C1), média-baixa (C2) e baixa (D). No entanto, poderão ser feitas análises onde entram outras variáveis sociodemográficas, como por exemplo, o género e idade.

² Os custos fixos são disponibilizados num ficheiro excel.

2. Apresentar o modelo multidimensional mais adequado a cada um dos processos de negócio;
3. Identificar os factos a reter para cada negócio;
4. Classificar cada um dos factos de acordo com a sua natureza;
5. Identificar a granularidade de cada uma das tabelas de factos;
6. Descrever o que representa cada tuplo em cada tabela de factos;
7. Identificar e incluir na definição das tabelas de dimensão, os atributos considerados relevantes tendo em conta a evolução futura do data warehouse;
8. Apresentar um esquema que ilustre o data warehouse bus.

1.2.Segunda Fase

1.2.1. Requisitos Mínimos

Utilizando a ferramenta SSIS, implementar o processo de ETL necessário para construir e actualizar o data warehouse descrito na primeira fase. Ter em particular atenção os possíveis problemas existentes na tabela Cliente e contemplar alterações na(s) dimensão(ões) que contenha essa informação.

1.2.2. Objectivos

1. Caracterizar um conjunto de dados com métricas relevantes ao desenvolvimento do ETL (data profiling);
2. Identificar e implementar soluções adequadas à resolução de inconsistências existentes nos dados provenientes do sistema OLTP;
3. Identificar e implementar as formas mais correctas de acesso aos dados do sistema OLTP, tendo em conta o desempenho global (OLTP+DW);
4. Projectar e implementar processos de transformação e carregamento de dados (ETL) que suportem um fluxo de transferência, temporizado, do sistema OLTP para o data warehouse contemplando actualizações de dados já transferidos e a inserção de novos dados;
5. Escolher a calendarização mais adequada para executar o processo ETL;
6. Utilizar a ferramenta SQL Server Integrations Services (SSIS) para implementar os processos de ETL, possibilitando a obtenção de configurações de repositórios externos ao pacote;

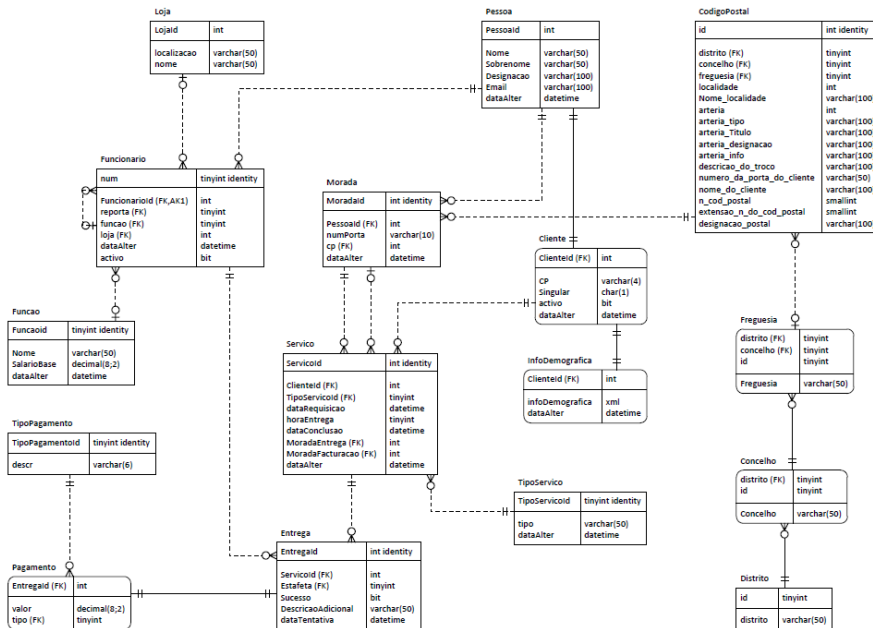
7. Identificar e implementar mecanismos de actualização dos atributos das tabelas de dimensão.

1.3. Notas para a Implementação

- a) Conceber os pacotes para serem idempotentes, ou seja, que possam ser executados várias vezes, sem que o resultado final seja alterado;
- b) Deixar espaço no pacote para uma última tarefa que iniciará o processamento do(s) cubo(s), se estes já existirem. Esta facilidade será explorada na fase seguinte.

2. Análise do Problema

2.1. Modelo Lógico do sistema Operacional



2.2. Classificação das entidades

A modelação multidimensional é utilizada para conceber a estrutura de sistemas de *Data Warehousing*. É baseada em dois pressupostos: produzir uma estrutura da base de dados fácil de compreender e utilizar, facilitando a colocação de interrogações ao sistema, e optimizar o desempenho no processo de questões, em oposição à optimização do processamento de actualizações, como se verifica no modelo relacional (que dá suporte ao OLTP).

As tabelas de factos constituem os componentes principais dos modelos multidimensionais, dado que permitem armazenar ou registar os acontecimentos a analisar. Estes acontecimentos estão associados aos seus respectivos processos de negócio. Na tabela de factos, um registo, ou linha da tabela, está associado a um dado acontecimento, devendo todos os acontecimentos estarem representados com a mesma granularidade dos dados. Esta granularidade representa o nível máximo de detalhe da informação armazenada.

Contudo há que analisar esta informação sobre diferentes perspetivas através de tabelas de dimensões que integram um conjunto diversificado de atributos, ou colunas, pelas quais os indicadores de negócio considerados nas tabelas de factos podem ser analisados. Estes atributos integram, habitualmente, descrições que permitem contextualizar as métricas em análise.

Commented [faf1]: Confuso: a rever

Para a identificação e caracterização destes tipos de tabelas, factos e dimensões, será utilizado o modelo de Moody, para a classificação de cada entidade contida no modelo EA nas seguintes topologias:

- a) Entidades transacionais;
- b) Entidades componentes;
- c) Entidade Classificativas;

3. Arquitectura da Solução

Para análise dos dados apresentados em uma estrutura multidimensional propõem-se a divisão da estrutura de negócio em dois processos, respetivamente:

- a) Análise Operacional;
- b) Análise Financeira.

A adaptação desta estratégia ao desenho da modelação dimensional, permite desenvolver pequenos Datamarts para cada um dos processos de negócio identificados, convergindo, a sua informação, posteriormente para o Datawarehouse da BIT, adotando-se neste caso o modelo proposto por Kimball.

Mode Pela análise dos dados de suporte ao OLTP separámos as entidades em três grupos, Entidades Transacionais, Entidades Classificativas e Entidades Componentes de que se junta um exemplo de mapeamento:

- a) Entidades transacionais (Serviço, Entregas);
- b) Entidades classificativas (TipoServiço, Loja, Função, CodigoPostal);
- c) Entidades componentes (Pessoa, Cliente, Morada, Funcionário);

3.1. Processo de Negócio "Operacional"

Este processo de negócio representa a atividade produtiva da empresa, ou seja os Serviços de Entrega das encomendas.

3.1.1. Identificação de Hierarquias

Para este processo foram identificadas as seguintes hierarquias:

- a) Pessoa->Morada->Serviço
- b) Pessoa->Morada->Serviço-> Entrega
- c) Distrito->Concelho->Freguesia -> CodigoPostal -> Morada -> Serviço
- d) Distrito->Concelho->Freguesia -> CodigoPostal -> Morada -> Serviço -> Entrega
- e) Pessoa -> Cliente -> Serviço
- f) Pessoa -> Cliente -> Serviço -> Entrega
- g) Pessoa -> Morada -> Serviço
- h) Pessoa -> Morada -> Serviço -> Entrega
- i) TipoServiço -> Serviço
- j) TipoServiço -> Serviço -> Entrega
- k) Loja -> Funcionario -> Entrega
- l) Loja -> Funcionario -> Entrega -> Serviço
- m) Pessoa -> Funcionario -> Entrega
- n) Pessoa -> Funcionario -> Entrega -> Serviço
- o) Função -> Funcionário -> Entrega
- p) Função -> Funcionário -> Entrega -> Serviço

3.1.2. Produção dos modelos multidimensionais

3.1.2.1. Contracção de Hierarquias

As tabelas de hierarquias do OLTP foram "desnormalizadas" de acordo com as Contracções indicadas a seguir:

- a) Morada←CodigoPostal←Freguesia←Concelho←Distrito
- b) Funcionario <- Chefe <- Pessoa <- Funcao <- Loja
- c) Cliente <- Pessoa
- d) Pagamento foi incluído na Entrega e no Serviço

3.1.2.2. Agregação

O processo de negócio da área Operacional é representado nas dimensões indicadas atrás e em duas tabelas de factos:

- Serviços – que representa para cada um dos “serviços” de entrega solicitados por cada um dos clientes. Tem uma relação de um para um com o sistema operacional (mas que não “limpa” ao contrário do que será expectável acontecer no OLTP). Agrupa indicadores como o número de tentativas e preenche colunas com o valor pago pelo serviço ou o sla do serviço;
- Entregas – representa cada uma das tentativas de entrega de um Serviço, tenha tido ou não sucesso. No caso de sucesso actualiza alguns atributos do Serviço respectivo.

No processo de passagem foram descartados alguns campos (como por exemplo o SalarioBase) por não serem necessários para os requisitos actuais e futuros (tanto quanto é possível prever neste momento) do data warehouse e das necessidades de pesquisa e análise dos utilizadores finais.

3.1.3. Data Data Profiling e Assunções

- a) Cliente
Nada a assinalar
- b) CodigoPostal
Na extensão do código postal assume-se que os valores inferiores a 100 serão preenchidos com zeros á esquerda;
- c) Concelho
Nada a assinalar
- d) DataLog
Nada a assinalar
- e) Distrito
Nada a assinalar
- f) Entrega
Existem casos de insucesso que não têm indicação do motivo;
- g) Freguesia
Nada a assinalar

- h) Função
Nada a assinalar
- i) Funcionário
Os funcionários gestores (Funcão=3) não têm hierarquia nem loja atribuída;
- j) InfoDemografica
Informação contida em estrutura XML, que precisa de ser convertida e integrada na dimensão de clientes
Nada a assinalar
- k) Loja
Nada a assinalar
- l) Morada
Existem clientes para os quais não está indicado o nº de porta;
- m) Pagamentos
Nada a assinalar
- n) Pessoa
designação = organizações/Nome+Sobrenome;
Existem empresas que são representados por diferentes "Nome+Sobrenome";
- o) Serviço
Nada a assinalar
- p) TipoPagamento
Nada a assinalar
- q) TipoServiço
Nada a assinalar

3.1.4. Arquitetura do Datamart "Operacional"

Para a análise dos dados apresentados em uma estrutura multidimensional, é definida uma arquitetura de um Datamart para armazenamento dos dados. Os dados carregados para este repositório serão posteriormente analisados recorrendo a uma ferramenta OLAP.

Na definição arquitetura, que conduzirá ao esquema em estrela do repositório, partimos dos seguintes pressupostos:

- a) O processo de negócio a analisar nesta vertente está associado a questões relacionadas com as entregas das encomendas;
- b) A granularidade adotada para o carregamento da informação disponível é ao nível da entrega por hora, venda, dia, cliente prevendo-se a agregação prévia da informação por dia;
- c) As dimensões que serão utilizadas para analisar o processo de negócio de entregas são:
 - i. Tipo_de_Serviço, Funcionario, Serviço, (Ano, Mês, Trimestre, Semana, Dia_da_Semana), Hora_do_Dia;
- d) O facto a analisar segundo as diferentes dimensões definidas no ponto anterior é o número de entregas, sendo que existirão dados derivados deste facto, tais como comissões baseadas nas vendas, quantidades;

Partindo destes pressupostos será possível então, efetuar as análises pretendidas que consistem na observação da evolução das entregas por tipo, loja, região, freguesia, cliente, ano, trimestre, mês, semana, dia da semana e hora.

Para este cenário, apresenta-se o esquema em estrela adotado para o Datamart para o processo de entregas da BIT. A opção por um esquema em estrela deriva do facto de termos identificado para análise apenas um processo. Deste modo o esquema em estrela integra a tabela de factos associada ao processo operacional ("ENTREGA"), bem como as dimensões sobre as quais os factos serão analisados.

Para análise de alguns atributos deste conjunto de dados recorrendo aos cubos OLAP, será necessário a definição e adoção de classes que permitam a agregação dos diversos valores disponíveis para o salário de cada cliente num conjunto mais restrito de descrições.

3.1.5. TABLE Operacoes.Morada

Tabela de dimensão, com a identificação do local de fim do serviço, resulta da composição de diversas outras tabelas do OLTP, é composta pelos seguintes campos:

keycol	INT PRIMARY KEY IDENTITY
oltp_id	INT UNIQUE,
rua	VARCHAR(302)
numeroDaPorta	VARCHAR(10)
nomeLocalidade	VARCHAR(100) NOT NULL
codigoPostal	INT NOT NUL,
extensaoCodigoPostal	INT
designacaoPostal	VARCHAR(100) NOT NULL

3.1.6. TABLE Operacoes.TipoServico

Tabela de dimensão, com tipificação dos tipos de serviços prestados pela BIT, é composta por:

keycol	TINYINT PRIMARY KEY IDENTITY
oltp_id	TINYINT UNIQUE NOT NULL
tipo	VARCHAR(50) NOT NULL

3.1.7. TABLE Operacoes.Cliente

Tabela de dimensão com a descrição de um cliente incluindo características infoDemográficas e cujos campos são:

keycol	INT PRIMARY KEY IDENTITY
oltp_id	INT UNIQUE NOT NULL

designacao	VARCHAR(100) NOT NULL
nomeContacto	VARCHAR(50) NOT NULL
sobrenomeContacto	VARCHAR(50) NOT NULL
singular	CHAR(1) NOT NULL
cPostal	VARCHAR(4) NOT NULL
activo	CHAR(1) NOT NULL
BirthDate	DATETIME NULL
MaritalStatus	VARCHAR(150) NULL
YearlyIncome	VARCHAR(150) NULL
Gender	VARCHAR(150) NULL
TotalChildren	VARCHAR(150) NULL
NumberChildrenAtHome	VARCHAR(150) NULL
Education	VARCHAR(150) NULL
Occupation	VARCHAR(150) NULL
HomeOwnerFlag	VARCHAR(150) NULL
NumberCarsOwned	VARCHAR(150) NULL

3.1.8. TABLE Operacoes.Funcionario

Tabela de dimensão com descrição de um funcionário incluindo elementos com a sua linha de reporting (directo):

Keycol	INT PRIMARY KEY IDENTITY
oltp_id	INT UNIQUE NOT NULL
funcao	VARCHAR(50) NOT NULL
nome	VARCHAR(50) NOT NULL,

sobrenome	VARCHAR(50) NOT NULL
activo	CHAR(1) NOT NULL
lojaNome	VARCHAR(50) NOT NULL
localizacao	VARCHAR(50) NOT NULL
chefeNome	VARCHAR(50)
chefeSobrenome	VARCHAR(50)
chefeFuncao	VARCHAR(50)
chefe_id	INT

3.1.9. TABLE Operacoes.Data

Tabela de dimensão Temporal, cujos campos são:

keycol	INT PRIMARY KEY,
data	DATE NOT NULL,
ano	INT NOT NULL,
trimestre	TINYINT NOT NULL,
mes	TINYINT NOT NULL,
descricaoMes	VARCHAR(15) NOT NULL,
diaSemana	TINYINT NOT NULL,
descricaoDiaSemana	VARCHAR(15) NOT NULL,
dia	TINYINT NOT NULL

3.1.10. TABLE Operacoes.Hora

Tabela de dimensão com as horas de um dia, para usar nas pesquisas e cujos campos são:

keycol	TINYINT PRIMARY KEY,
hora24	VARCHAR(5) NOT NULL,
hora12	VARCHAR(8) NOT NULL

3.1.11. TABLE MonthDescr

Tabela (auxiliar) de dimensão com a descrição do nome dos meses em Português.

keycol	TINYINT PRIMARY KEY,
nome	VARCHAR(15) NOT NULL

3.1.12. TABLE DiaSemanaDescr

Tabela (auxiliar) de dimensão com a descrição do nome dos dias da semana em Português.

keycol	TINYINT PRIMARY KEY,
nome	VARCHAR(15) NOT NULL

3.1.13. TABLE Operacoes.Servicos

Tabela de factos que descreve um Serviço, ou seja uma das operações de valor da BIT e que tem os seguintes campos:

keycol	INT PRIMARY KEY IDENTITY,
oltp_id	INT UNIQUE NOT NULL,
cliente	INT REFERENCES Operacoes.Cliente,
moradaEntrega	INT REFERENCES Operacoes.Morada,
moradaFacturacao	INT REFERENCES Operacoes.Morada,
tipoServico	TINYINT REFERENCES Operacoes.TipoServico,
dataRequisicao	INT REFERENCES Operacoes.Data,
horaRequisicao	TINYINT REFERENCES Operacoes.Hora,
dataConclusao	INT REFERENCES Operacoes.Data,
horaConclusao	TINYINT REFERENCES Operacoes.Hora,
horaEntrega	TINYINT REFERENCES Operacoes.Hora,
numeroEntregas	INT NOT NULL, -- count(Entregas)
valor	Decimal(8,2),
dentroDoSLA	TINYINT,
concluido	TINYINT NOT NULL

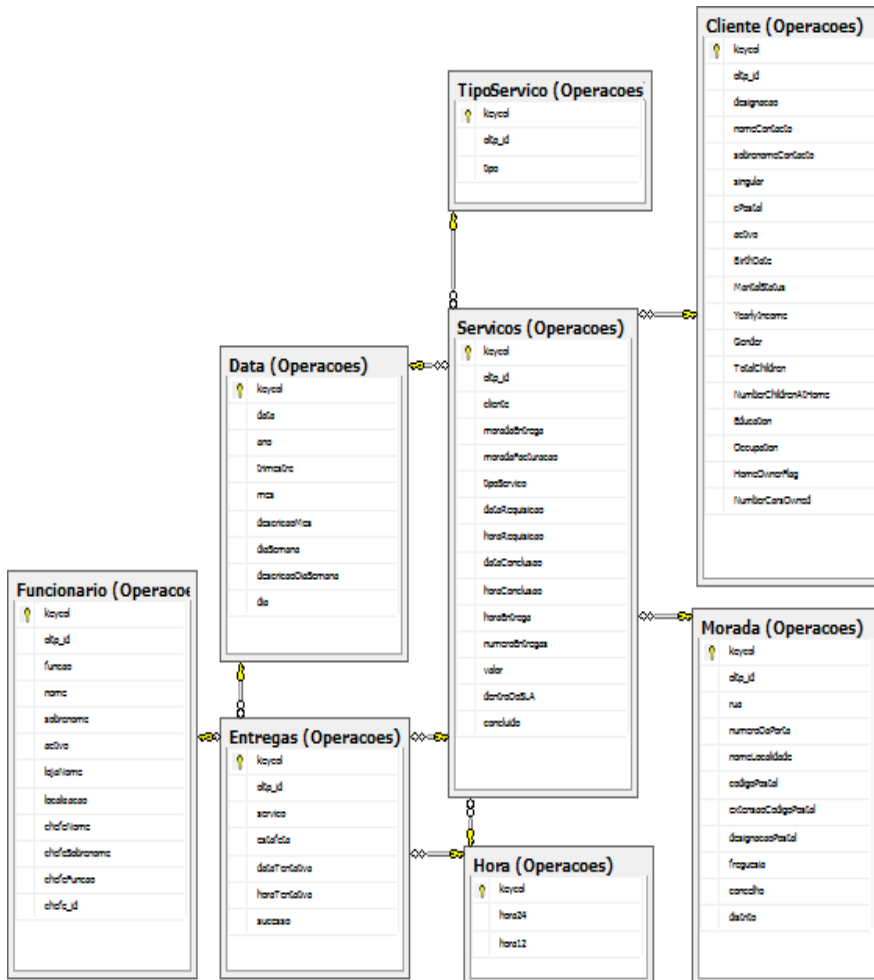
Os campos dentroDoSLA e concluído, são campos que descrevem um evento binário e para os quais é aqui usado um 1 para permitir a soma de valores igual à contagem.

3.1.14. TABLE Operacoes.Entregas

Tabela de factos, que descreve cada uma das tentativas de “entrega” de uma encomenda de um dado serviço e que poderá ser feita com sucesso, ou não.

keycol	INT PRIMARY KEY IDENTITY,
oltp_id	INT UNIQUE NOT NULL,
servico	INT REFERENCES Operacoes.Servicos,
estafeta	INT REFERENCES Operacoes.Funcionario,
dataTentativa	INT REFERENCES Operacoes.Data,
horaTentativa	TINYINT REFERENCES Operacoes.Hora,
sucesso	TINYINT NOT NULL

3.1.15. Modelo do DATAMART Operacional



3.1. Processo de Negócio “Financeiro”

Para a análise dos dados apresentados numa estrutura multidimensional, é definida uma arquitetura de um Datamart para armazenamento dos dados. Os dados carregados para este repositório serão posteriormente analisados recorrendo a uma ferramenta OLAP.

Na definição arquitetura, que conduzirá ao esquema em estrela do repositório, partimos dos seguintes pressupostos:

- e) O processo de negócio a analisar nesta vertente está associado a questões relacionadas com a gestão financeira da empresa (informação sobre o balanço mensal);
- f) A granularidade adotada para o carregamento da informação disponível é mensal para cada uma das rubricas consideradas (no balanço):
 - i. Proveitos: Serviços;
 - ii. Custos: Materiais, Financeiros, Salários e Outros ao nível dos custos e proveitos mensais tendo por base a informação recebida em excel ou agregada no OLTP ou “Área Operacional” de acordo com o grão referido;
- g) As dimensões que serão utilizadas para analisar o processo de negócio relativo ao balanço mensal são:
 - i. Ano, Trimestre e Mês
 - ii. Proveito ou Custo
 - iii. Rubrica (do balanço)
- h) O facto a analisar segundo as diferentes dimensões definidas no ponto anterior os resultados financeiros mensais, sendo que poderão existir dados derivados deste facto;

Partindo destes pressupostos será possível então, efetuar as análises pretendidas que consistem na observação da evolução dos resultados financeiros por tipo de custo, ano, trimestre, e mês.

Para este cenário, apresenta-se o esquema em estrela adotado para o Datamart para o processo de balanço da BIT. A opção por um esquema em estrela deriva do facto de termos identificado para análise apenas um processo. Deste modo o esquema em estrela

integra a tabela de factos associada ao processo financeiro ("BALANÇO"), bem como as dimensões sobre as quais os factos serão analisados.

Para análise de alguns atributos deste conjunto de dados recorrendo aos cubos OLAP, será necessário a definição e adoção de classes que permitam a agregação dos diversos valores disponíveis para o salário de cada cliente num conjunto mais restrito de descrições.

3.1.1. TABLE Operacoes.Meses

Tabela de dimensão temporal com os seguintes campos:

keycol	INT PRIMARY KEY IDENTITY
ano	INT NOT NULL
trimestre	TINYINT NOT NULL
mes	TINYINT NOT NULL
descricaoMes	VARCHAR(15) NOT NULL

3.1.2. TABLE Operacoes.Rubrica

Tabela de dimensão com as rubricas do balanço agrupadas em duas classes: proveitos e custos. Tendo os proveitos um valor positivo e os custos um valor negativo, é possível de forma simples obter o "resultado operacional" para o conjunto selecionado de custos.

keycol	INT PRIMARY KEY IDENTITY
classe	VARCHAR(15) NOT NULL
nome	VARCHAR(50) NOT NULL

Nota: o campo nome identifica as rubricas consideradas e a classe o seu impacto no resultado. Os valores criados foram:

- Proveito:
 - Serviços – somatórios do valores recebidos pelos serviços (entrega de encomendas) de acordo com o mês da sua conclusão (um serviço poderá ter várias tentativas de entrega, a data de conclusão é a do último);

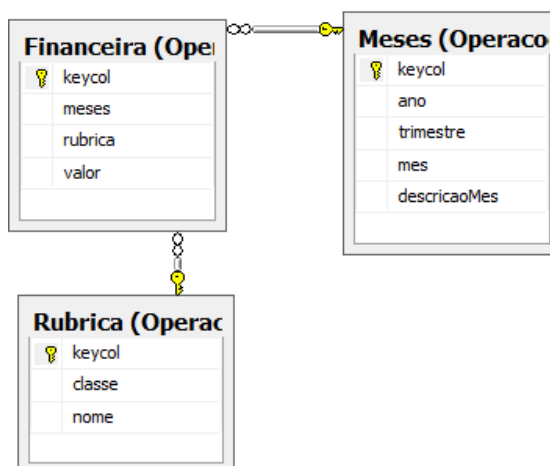
- Custos:
 - o Materiais - informação recolhida do excel de custos;
 - o Financeiros - informação recolhida do excel de custos;
 - o Salários - custos de todos os funcionários activos ou até ao mês da data de alteração (assumindo que a última alteração foi a desactivação do funcionário e considerando meses completos);
 - o Outros - informação recolhida do excel de custos.

3.1.3. TABLE Operacoes.Financeira

Tabela de factos financeiros, apresenta uma linha mensal para cada rubrica considerada (ver descrição da dimensão Rubrica)

keycol	INT PRIMARY KEY IDENTITY
meses	INT REFERENCES Operacoes.Meses
rubrica	INT REFERENCES Operacoes.Rubrica
valor	DECIMAL(10,2) NOT NULL

3.1.4. Modelo do DATAMART Financeiro



4. Desenvolvimento e Implementação do ETL

O ETL (Extract Transform Load) foi desenvolvido com suporte no SSIS (SQL Server Integration Services), através de respectivo projecto do Visual Studio. O SSIS permite a criação de soluções para transformação e integração de dados em grande escala. Nomeadamente na integração entre o sistema OLTP (BIT), e o sistema OLAP (DataStagingBIT) e respectivos Datamarts (Operacional e Financeiro).

4.1. Descrição da Integração

No projecto com.isel.siad.sv1213.bit.ssis, encontra-se o package principal do ETL, BitMainPackage.dtsx, composto pelas seguintes tarefas:

1. Criar Datastaging (Execute SQL Task)

Execução de script sql, para criação da base de dados DataStagingBIT, e todos os objectos correspondentes aos modelos multidimensionais dos dois processos de negócio.

2. Execute OperationalPackage (Execute PackageTask)

Execução de outro package, OperationalPackage.dtsx, responsável pela transformação e integração de dados, com origem no sistema OLTP BIT, e destino o Datamart “Operacional”.

3. Copy remote XLS (Script Task)

Código C#.net que faz o download do ficheiro custos.xls, e guarda em directoria local, para que possa ser utilizado no package financeiro.

4. Execute FinacialPackage (Execute PackageTask)

Execução de outro package, OperationalPackage.dtsx, responsável pela transformação e integração de dados, com as seguintes origens:

- Sistema OLTP Bit
- Ficheiro custos.xls
- Datamart “Operacional”, já carregado na tarefa 2

E tem como destino o Datamart “Financeiro”.

4.2. Anexos

Em anexo segue a pasta da solução (com.isel.siad.sv1213.bit), contendo tudo o que é necessário para testar o ETL desenvolvido.

4.2.1. Configuração

Antes de abrir a solucao com.isel.siad.sv1213.bit.sln, efectuar as seguintes configuracoes:

Considerando que (path) é o caminho para a pasta onde se encontra a pasta de repositório local "com.isel.siad.sv1213.bit".

No ficheiro:

(path)\com.isel.siad.sv1213.bit\com.isel.siad.sv1213.bit.ssis\BitMainPackage.Variables.dtsConfig

Substituir directoria:

D:\ISEL@Local\SIAD\com.isel.siad.sv1213.bit\com.isel.siad.sv1213.bit.ssis\FinacialPackage.dtsx

Por directoria:

(path)\com.isel.siad.sv1213.bit\com.isel.siad.sv1213.bit.ssis\FinacialPackage.dtsx

Substituir directoria:

D:\ISEL@Local\SIAD\com.isel.siad.sv1213.bit\#SQL\ETL\

Por directoria:

(path)\com.isel.siad.sv1213.bit\#SQL\ETL\

Substituir directoria:

D:\ISEL@Local\SIAD\com.isel.siad.sv1213.bit\com.isel.siad.sv1213.bit.ssis\OperationalPackage.dtsx

Por directoria:

(path)\com.isel.siad.sv1213.bit\#SQL\ETL\

No ficheiro

(path)\com.isel.siad.sv1213.bit\com.isel.siad.sv1213.bit.ssis\BitMainPackage.Connections.dtsConfig

Substituir Connection String:

```
Data Source=10.137.41.4\SQLEXPRESS;User ID=siad_bit_dev;  
Password=siad_bit_dev;Provider=SQLNCLI10.1;Persist Security Info=True;
```

Pela Connection String que aponta para o servidor onde ficará a base de dados de DataStaging.

Nos ficheiros:

(path)\com.isel.siad.sv1213.bit\com.isel.siad.sv1213.bit.ssis\OperationalPackage.Connections.dtsConfig

(path)\com.isel.siad.sv1213.bit\com.isel.siad.sv1213.bit.ssis\FinancialPackage.Connections.dtsConfig

Substituir Connection String:

```
Data Source=10.137.41.4\SQLEXPRESS;User ID=siad_bit_dev; Password=siad_bit_dev;Initial Catalog=DataStagingBIT;Provider=SQLNCLI10.1;Persist  
Security Info=True;
```

Pela Connection String que aponta para o servidor onde se encontra a base de dados de DataStaging.

Ao abrir solução "com.isel.siad.sv1213.bit.sln" e durante a utilização, se for pedida alguma password colocar "siad_bit_dev" (sem aspas).

Executar pacote BitMainPackage.dtsx (este pacote chamará os outros pacotes).

5. Considerações Finais:

5.1. Geral

Sendo um modelo inicial é previsível que durante a fase inicial do processo de utilização, seja necessário proceder a pequenos reajustamentos, quer na tabela de factos quer nas tabelas de dimensões.

As tabelas de dimensões integram um conjunto diversificado de atributos, ou colunas, pelos quais os indicadores de negócio considerados nas tabelas de factos poderão ser avaliados e analisados. Estes atributos integram, habitualmente descrições que permitem contextualizar as métricas em análise.

5.2. Melhorias

Neste processo de aprendizagem de implementação do modelo Dimensional, tomámos algumas decisões, que fomos repensando durante o projeto e que são candidatos a uma correcção ou melhoria. Este neste caso os seguintes:

- Processo de carregamento da dimensão Morada, está lento e deveria ser revisto para identificar formas de melhorar o desempenho;
- Geração dos valores de custos de Salários, está a ser usado um join com uma tabela do OLTP (os dados em causa não existem no DW), pensamos que terá o mesmo impacto que a sua extração pura e simples, mas talvez tenha de ser revisto para permitir historico e carregamentos incrementais;
- Separar os Processos de Negócio por "schemas", na implementação usámos apenas um "schema" – Operacoes, o que não interfere nos carregamentos e na actividade desta primeira fase, mas que permitirá simplificar o uso futuro do sistema;
- Seria interessante dispôr de uma mais fácil parametrização dos packages de ETL sem ter de usar a tool de desenvolvimento, contudo a investigação da funcionalidade pode ser morosa, pelo que optamos por a deixar para uma próxima versão;
- O package ETL para a área financeira, não permite uma execução parcial repetida, embora permita dentro do processo total (com o ETL para a área de Operações), a

separação dos Schemas e a técnica usada em “Operações” resolve este issue, que não endereçamos agora por uma questão de tempo (implementação + testes);

5.3. Futuro

Dado que o Data Warehouse é um repositório de leitura de dados e que as operações de escrita desses dados está restringida ao carregamento e “refrescamento” do Data Warehouse, não é possível, fora deste contexto, a alteração dos valores contidos nas diversas dimensões.

As alterações aos valores contidos nas diversas dimensões só poderão ser efetuadas nos períodos em que existe atualizações ao Data Warehouse. No entanto, será necessário definir previamente qual a estratégia a adotar para a realização de tais atualizações. De salientar que este tipo de alterações não são muito frequentes, e por este motivo, são designadas por “slowly changing dimension”, podendo ser observadas três possibilidades de alterações:

Tipo 1 – Alteração de Valor (substituição do tupulo);

Tipo 2 – Inserção de novo tupulo;

Tipo 3 – Combinação dos anteriores.

A inserção de novos tupulos nas tabelas de dimensões é a alternativa de utilização mais frequente, uma vez que permite preservar todo histórico de todas as alterações efetuadas na tabela, sendo de todas as hipóteses, a de maior complexidade de implementação.

Para tal efeito, uma das técnicas mais utilizadas, consiste na utilização de uma chave primária composta, normalmente com dois componentes. O primeiro permite identificar determinado registo, o segundo identifica as diversas alterações introduzidas no tupulo, sendo deste modo designadas como chaves de estrutura. Um das hipóteses de utilização poderá ser a alteração de morada de um cliente, a mudança de agente de uma banda, etc.

Haverá, que considerar também a hipótese de prever alterações pela criação prévia de colunas para alterações, mas cuja eficácia é discutível, sendo necessário prever algumas colunas adicionais para conter as referidas alterações.