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:** | **2013/03/09** |



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** |  |
|  |  | |  |  |  |
|  |  | |  |  |  |

Índice

[2. Síntese do Caso 3](#_Toc353029928)

[2.1. Primeira Fase 3](#_Toc353029929)

[2.1.1. Requisitos Mínimos 3](#_Toc353029930)

[2.1.2. Objectivos 4](#_Toc353029931)

[2.2. Segunda Fase 4](#_Toc353029932)

[2.2.1. Requisitos Mínimos 4](#_Toc353029933)

[2.2.2. Objectivos 4](#_Toc353029934)

[2.3. Notas para a Implementação 5](#_Toc353029935)

[3. Desenvolvimento e Implementação 6](#_Toc353029936)

[3.1. Primeira Fase 6](#_Toc353029937)

[3.1.1. Modelo Lógico do sistema Operacional 6](#_Toc353029938)

[3.1.2. Classificação das entidades 6](#_Toc353029939)

[3.1.3. Identificação de Hierarquias 8](#_Toc353029940)

[3.1.4. Desenvolvimento dos modelos multidimensionais 8](#_Toc353029941)

[3.1.5. Data Data Profiling e Assunções 8](#_Toc353029942)

# 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, tácticas e estratégicas, a empresa decidiu investir na melhoria desse sistema, através da implementação de um data warehouse, que integra toda a informação.

## Primeira Fase

### 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, cliente[[1]](#footnote-1), 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 fixos[[2]](#footnote-2) e o valor dos pagamentos, pretendendo-se efectuar análises por mês, trimestre, ano, tipo de custo.

### Objectivos

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

## Segunda Fase

### 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.

### 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.

## Notas para a Implementação

1. Conceber os pacotes para serem idempotentes, ou seja, que possam ser executados várias vezes, sem que o resultado final seja alterado;
2. 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.

# Desenvolvimento e Implementação

## Primeira Fase

Desenho – Modelo de Kimball / modelo de Moody ()

### Modelo Lógico do sistema Operacional



### 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.

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. Numa tabela de factos, um registo, ou linha da tabela, está associado a um, dado acontecimento, devendo todos os acontecimentos estarem representados recorrendo á mesma granularidade dos dados. Esta granularidade representa o nível de detalhe da informação armazenada.

Contudo há que analisar esta informação sobre diferentes perspectivas 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.

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:

1. Entidades transaccionais;
2. Entidades componentes;
3. Entidade Classificativas;
4. Entidades transaccionais (Serviço, Entregas);
5. Entidades componentes (TipoServiço, Cliente, Morada, Funcionário):
6. Entidade Classificativas (Pessoa, Loja, Função, CodigoPostal);

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

1. Análise Operacional;
2. Análise Financeiro.

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.

### Identificação de Hierarquias

1. Servico→Morada→Pesssoa
2. Servico→Morada→CodigoPostal→Freguesia→Concelho→Distrito
3. Entrega→Servico→Cliente→Pessoa
4. Entrega→Servico→Morada→Pesssoa
5. Entrega→Servico→TipoServico
6. Servico→Morada→CodigoPostal→Freguesia→Concelho→Distrito
7. Servico→Cliente→Pessoa
8. Entrega→Funcionario→Loja
9. Entrega→Funcionario→Pessoa
10. Entrega→Funcionario→Função

### Desenvolvimento dos modelos multidimensionais

#### Contracção de Hierarquias

1. Morada←CodigoPostal←Freguesia←Concelho←Distrito

#### Agregação

### Data Data Profiling e Assunções

1. Cliente

Nada a assinalar -

1. CodigoPostal

Na extensão do código postal assume-se que os valores inferiores a 100 serão preenchidos com zeros á esquerda;

1. Concelho
2. Nada a assinalar -
3. DataLog
4. Nada a assinalar -
5. Distrito

Nada a assinalar -

1. Entrega

Existem casos de insucesso que não têm indicação do motivo;

1. Freguesia
2. Nada a assinalar -
3. Função

Nada a assinalar -

1. Funcionário

Os funcionários gestores (Funcão=3) não têm hierarquia nem loja atribuída;

1. InfoDemografica
2. Nada a assinalar -
3. Loja

Nada a assinalar -

1. Morada

Existem clientes para os quais não está indicado o nº de porta;

1. Pagamentos

Nada a assinalar -

1. Pessoa

designação = organizações/Nome+Sobrenome;

Existem empresas que são representados por diferentes “Nome+Sobrenome”;

1. Serviço

Nada a assinalar -

1. TipoPagamento

Nada a assinalar -

1. TipoServiço

Nada a assinalar -

# Arquitetura Geral

## Arquitetura do DATAMART - “Análise\_Operacional”

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:

1. O processo de negócio a analisar nesta vertente está associado a questões relacionadas com as entregas das encomendas;
2. 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;
3. As dimensões que serão utilizadas para analisar o processo de negócio de entregas são:
   1. Tipo\_de\_Serviço, Loja, Região, Freguesia, Cliente, (Ano, Mês, Trimestre, Semana, Dia\_da\_Semana), Hora\_do\_Dia;
4. 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 prossupostos 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.

|  |  |  |
| --- | --- | --- |
| **Tipo Tabela** | **Tabela** | **Colunas** |
| Dimensões | **Tipo\_de\_Serviço** | Keycol INT PRIMARY KEY IDENTITY,  oltp\_id TINYINT UNIQUE NOT NULL,  Tipo\_de\_Serviço NVARCHAR(20) NOT NULL |
|  | |
| **DateStamp** | Keycol INT PRIMARY KEY IDENTITY  Data DATE NOT NULL  Ano INT NOT NULL  Trimestre TINYINT NOT NULL  Mes TINYINT NOT NULL  Nome\_Mes NVARCHAR(15) NOT NULL  Dia TINYINT NOT NULL  Dia\_da\_Semana TINYINT NOT NULL  Nome\_Dia\_da\_Semana NVARCHAR(15) NOT NULL |
|  | |
| **Hora\_do\_Dia** | keycol INT PRIMARY KEY IDENTITY  oltp\_id TINYINT UNIQUE NOT NULL  Hora\_do\_Dia NVARCHAR(10) NOT NULL, |
|  | |
| **Loja** | keycol INT PRIMARY KEY IDENTITY  oltp\_id TINYINT UNIQUE NOT NULL Nome\_da\_Loja NVARCHAR(15) NOT NULL |
|  | |
| **Morada** | keycol INT PRIMARY KEY IDENTITY  oltp\_id INT UNIQUE NOT NULL  Rua NVARCHAR(100) NOT NULL  Localidade NVARCHAR(100) NOT NULL  Codigo\_Postal NVARCHAR(8) NOT NULL  Designacão\_Postal NVARCHAR(100) NOT NULL  Freguesia NVARCHAR(100) NOT NULL  Concelho NVARCHAR(100) NOT NULL  Distrito NVARCHAR(100) NOT NULL |
|  | |
| **Pessoa** | keycol INT PRIMARY KEY IDENTITY  oltp\_id INT UNIQUE NOT NULL  Nome\_Cliente NVARCHAR(100) NOT NULL  Status\_Cliente NVARCHAR(10) NOT NULL |
|  | |

|  |  |  |
| --- | --- | --- |
| Factos | **Serviços\_de\_Entrega** | keycol INT PRIMARY KEY IDENTITY  Loja INT REFERENCES  Nome\_Cliente INT REFERENCES  Nome\_Estafeta INT REFERENCES  Tipo\_de\_Servico INT REFERENCES  Morada\_de\_Entrega INT REFERENCES  Data\_Requisicao\_Serviço INT REFERENCES  Hora\_Entrega\_Serviço INT REFERENCES  Data\_Execucao\_Serviço INT REFERENCES  Status\_Entrega INT REFERENCES  Modo\_de\_Pagamento INT REFERENCES  Valor\_Serviço decimal(8,2) NOT NULL |
| Criação da Tabela de factos "Entregas" que regista as entregas realizadas por cada estafeta.  **Nota:** *Para além das chaves associadas às Tabelas de Dimensão foi associada à Chave Primária o identificador de Entrega de modo a identificar univocamente uma entrega. Existem vários cálculos que deverão ser realizados e registados nesta Tabela* | |

## Arquitetura do DATAMART - “Análise\_Finaneira”

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:

1. O processo de negócio a analisar nesta vertente está associado a questões relacionadas com a gestão financeira da empresa (informação sobre o balanço mensal);
2. A granularidade adotada para o carregamento da informação disponível é ao nível dos custos e proveitos mensais prevendo-se a agregação prévia da informação por mês;
3. As dimensões que serão utilizadas para analisar o processo de negócio relativo ao balanço mensal são:
   1. Tipo\_Custo, (Ano, Trimestre, Mês);
4. 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 prossupostos 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.

|  |  |  |
| --- | --- | --- |
| **Tipo Tabela** | **Tabela** | **Colunas** |
| Dimensões | **Tipo\_de\_Custo** | Keycol INT PRIMARY KEY IDENTITY,  oltp\_id TINYINT UNIQUE NOT NULL,  Tipo\_de\_Custo NVARCHAR(20) NOT NULL |
|  | |
| **DateStamp** | Keycol INT PRIMARY KEY IDENTITY  Data DATE NOT NULL  Ano INT NOT NULL  Trimestre TINYINT NOT NULL  Mes TINYINT NOT NULL  Nome\_Mes NVARCHAR(15) NOT NULL |
|  | |
|  | keycol INT PRIMARY KEY IDENTITY  oltp\_id INT UNIQUE NOT NULL  XX NVARCHAR(10) NOT NULL  XX NVARCHAR(10) NOT NULL |
|  | |

|  |  |  |
| --- | --- | --- |
| Factos | **Dados\_Financieros** | keycol INT PRIMARY KEY IDENTITY  Tipo\_de\_Custo INT REFERENCES  Valor\_Serviço decimal(8,2) NOT NULL |
| Criação da Tabela de factos "Entregas" que regista os valores mensais( Custos/Proveitos) da BIT.  **Nota:** *Para além das chaves associadas às Tabelas de Dimensão foi associada à Chave Primária o identificador de Entrega de modo a identificar univocamente uma entrega. Existem vários cálculos que deverão ser realizados e registados nesta Tabela.* | |

# Considerações Finais:

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.

Dado que o DataWarehouse é um repositório de leitura de dados e que as operações de escrita desses dados está restringida ao carregamento e “refrescamento” do DataWarehouse, 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 DataWarehouse. 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.

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. [↑](#footnote-ref-1)
2. Os custos fixos são disponibilizados num ficheiro excel. [↑](#footnote-ref-2)