UNIVERSIDADE ABERTA
www.uab.pt

Projeto Final em Engenharia Informática

Projeto de Data Warehouse

IMPLEMENTAÇÃO E ANÁLISE

Pedro Miguel Sequeira Narciso

Professor Luís Cavique

Lisboa 5 de Novembro de 2020

Conteúdo

Lista de figuras 3
Lista de tabelas 4
Lista de abreviaturas 5
Introdução 6
Organização do relatório 6
Cronograma 6
Capítulo 1 – Levantamento bibliográfico 7
1.1. Data warehouse 7
1.2. William (Bill) H. Inmon 9
1.3. Ralph Kimball 11
1.4. Implementação top-down versus bottom-up 13
1.5. ETL - Extract Transform Load 14
1.6. Structed Query Language - SQL 16
1.7. Cubos e olap 17
Capítulo 2 – Data Sources (Fontes dos dados) 20
2.1. Referência para o Estudo de caso 20
2.2. Estrutura de dados 20
Capítulo 3 – Matriz de Oferta-procura da Informação 25
3.1 – Narrativa do SI 25
3.2- Matriz Oferta-Procura 25
3.3- ETL 25
3.4- Perguntas OLAP (5 perguntas) 25
Capítulo 4 - Conclusão 26
Bibliografia 27

Lista de figuras

Figura 1: ERD departamental. (Inmon, 2002) 10
Figura 2: O DIS da empresa. (Inmon, 2002) 10
Figura 3: Construtores de segundo nível de modelação (Inmon, 2002) 11
Figura 4:Tabelas de facto e dimensões num modelo dimensional (Kimball & Ross, 2002) 12
Figura 5:O processo de ETL – Fonte: (Kimball, Ralph; Caserta, Joe, 2004) 15
Figura 6: Star Schema versus cubo OLAP - fonte: (Kimball & Ross, 2002) 19
Figura 7: sakila Schema, (fonte própria) 21
Figura 8: Schema emplyees do departamento recursos humanos. 23
Figura 9: Schema suppliers do departamento recursos aquisições..... 24

Lista de tabelas

Tabela 1: Tipo de tabelas no schema sakila 21
Tabela 2 : Tabela de factos do schema sakila 22
Tabela 3: Tipo de tabelas no schema employee 23
Tabela 4: Tabela de factos do schema employee 23
Tabela 5: Tipo de tabelas no schema suppliers 24
Tabela 6: Tabela de factos do schema suppliers 25

Lista de abreviaturas

DM – Data Marts

DW – Data Warehouse

DWB – Data Warehouse Bus

ETL - Extract Transform and Load

OLAP – Online analytic processing

OLTP – Online transaction processing

RDBMS - Relational Database Management System

SQL – Structed Query Language

Introdução

Este trabalho foca-se numa implementação eficaz de uma data warehouse (DW) a qual será elaborada com foco na base de dados de exemplo *sakila*. (Oracle, 2020) Para implementação da DW irá proceder-se ao estudo de uma empresa fictícia, designada por VideoX cuja área de atividade é o aluguer de vídeos em lojas físicas. A VideoX possui nos sistemas informáticos três bases de dados operacionais: A base de dados *sakila*, dedicada ao foco negocial da empresa, uma base de dados de dedicada aos recursos humanos, e uma base de dados de fornecedores na qual se regista as compras e os contactos efetuados.

Ao longo do trabalho irá proceder-se à analise das fontes de dados disponíveis, e recorrendo a casos de estudo desenvolver uma DW com foco na análise das alternativas de implementação existentes.

ORGANIZAÇÃO DO RELATÓRIO

O trabalho está organizado por temas que visam descrever o processo de desenvolvimento. Um primeiro tema será dedicado ao levantamento bibliográfico, segue-se uma análise da empresa VideoX, com o levantamento de requisitos e em último os casos de estudo para implementar com eficácia a DW.

Neste relatório incidirá principalmente sobre o levantamento bibliográfico, e análise da empresa VideoX, ficando a parte de desenvolvimento e implementação para o relatório final.

CRONOGRAMA

Capítulo 1 – Levantamento bibliográfico

Revisão da bibliografia relevante nos temas data warehouse e de extração tratamento e carregamento de dados, ETL na sigla em Inglês para *Extract Transform and Load*.

1.1. DATA WAREHOUSE

Definições e propósito de uma *data warehouse*

> *"A collection of integrated, subject-oriented databases designed to supply the information required for decision-making."* (Inmon, 2002)

O termo *data warehouse* foi primeiramente apresentado por Bill Inmon no livro *Building the Data Warehouse. 1st Edition. Wiley and Sons* publicado em 1992. Este livro e as edições subsequentes foram em conjunto com a série lançada por Ralph Kimball *The Data Warehouse Toolkit, Willey and Sons* publicado em 1996 os percursores de duas metodologias de implementação de uma data warehouse.

Data warehouse não é considerada uma nova tecnologia, mas sim uma nova forma de aproveitar as tecnologias existes para de forma concertada resolver um problema que a maior parte das empresas e instituições vinham a sentir relativamente aos dados que vinham acumulando nos seus sistemas. A necessidade de recolha de informação em sistemas de bases de dados operacionais levantava muitas dificuldades, e eram baseadas em relatórios extraídos segundo visões diferentes daqueles a quem eram solicitados. Não raras vezes relatórios solicitados a diferentes departamentos resultavam em informações dispares entre eles.

O conceito de *data warehouse* advém da necessidade de existir uma ferramenta dedicada ao suporte da decisão. As bases de dados tradicionais apresentam como principal caraterística o facto de serem estruturadas para transações e customizados para sistemas integrados, com o principal foco nos processos e requerimentos de determinado departamento dentro de uma empresa. Numa *data warehouse* pretende-se ir além das visualizações tradicionais, concentrando-se nos assuntos

que abrangem toda a empresa, fornecendo um quadro único e completo dos processos de negocio.

Bill Inmon e Ralph Kimball, pioneiros e visionários em data warehouse, criaram conceitos e princípios de modelagem que ainda hoje são vistos com referências. Mesmo a trabalharem de forma independente e com modelos distintos os conceitos e princípios apresentados possuem muitos pontos em comum. Fon Silvers [SILVERS 2008] descreve-esses conceitos e princípios da seguinte forma:

Orientado ao assunto

Os dados numa DW não refletem as transações operacionais, mas devem apresentar-se como um reflexo das áreas temáticas da empresa. Estes devem ser organizados de maneira a que reflitam os assuntos e de forma a que seja possível manter a sua arquitetura ao longo do tempo.

Integração de Dados

Numa data warehouse (DW) os dados devem ser uniformes de forma a permitir a sua exploração por diversas áreas de forma coerente e com o mesmo significado.

Não volatilidade

Os dados uma vez carregados em DW devem ser persistentes e não estarem sujeitos a atualizações, e não devem ser apagados destes sistemas.

Dados temporais

Os dados devem ser expressos com uma referência temporal de forma a expressar os eventos empresariais numa perspetiva e num contexto histórico.

Uma única versão dos dados

Os dados devem ser moldados e guardados de forma a apresentar uma única versão da realidade, e a sua origem deve de ser a versão única e não ter origem em cópias que por vezes são efetuadas para reaproveitamento. Restringe-se o seu valor a apenas uma realidade de forma a que só um tipo de resposta possa ser retribuído sobre essa realidade.

Investimento a longo termo

Porque se pretende que uma DW seja um registo histórico é importante que os dados tenham longevidade e estabilidade. Uma DW deve ser flexível o suficiente para puder acomodar o crescimento da empresa e crescer com a variante histórica.

1.2. WILLIAM (BILL) H. INMON

Bill Inmon (Inmon, 2002) apresenta um modelo de implementação de uma data warehouse segundo uma metodologia que segue os princípios de uma visão *top down*. Este método surge como uma evolução natural do modelo relacional ficando conhecido com *Third normal form data model* o qual permite refletir os dados representados nos DBMS's da empresa.

Metodologia

O desenvolvimento da data warehouse deve ser efetuado com base em 3 níveis de modelação dos dados. Num primeiro nível a que Bill Inmom designa pelo mais elevado as diferentes áreas de negócio da empresa ou departamentos devem definir por si os dados relevantes e apresentar um modelo de diagrama entidade-relação, (ERD - Entity Relation Diagram) o qual deve representar as entidades, os atributos e relacionamentos dentro dos dados disponíveis nas bases de dados que lhes são afetas. Estes modelos devem manter-se o mais fiáveis relativamente ao modelo relacional das bases de dados de origem., com as adaptações necessárias mantendo as caraterísticas relacionais. Do resultado da concatenação dos diferentes modelos de dados resultantes é criado um modelo empresarial.

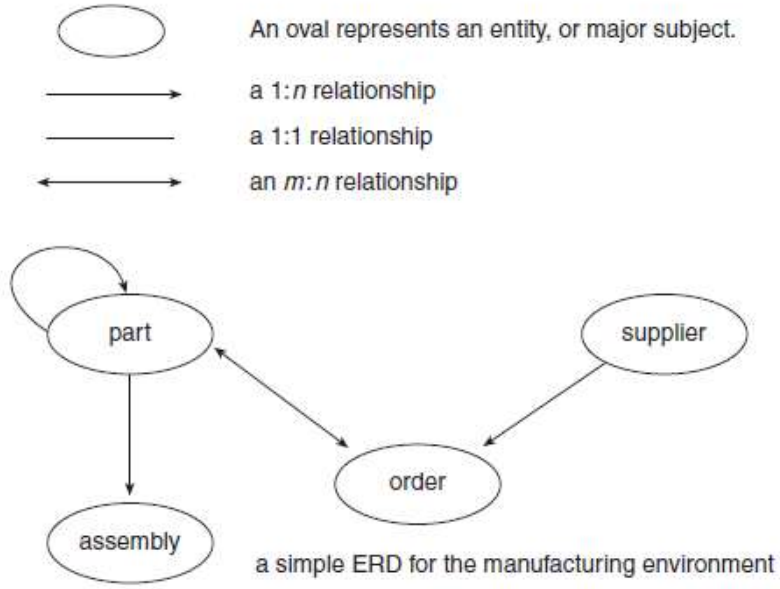

Figura 1: ERD departamental. (Inmon, 2002)

Num segundo nível de modelação designado por nível central, em que se estabelece o conjunto de dados (DIS – Data Item Set) para cada entidade do modelo ERD.

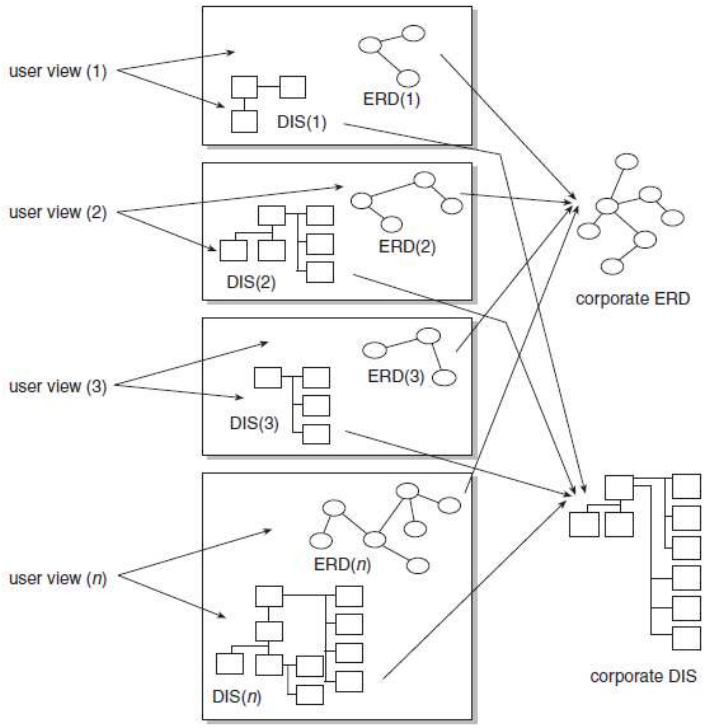

Figura 2: O DIS da empresa. (Inmon, 2002)

Os vários DIS são integrados através de 4 construtores sequenciais, em que o primeiro efetua um agrupamento dos dados nos atributos que não se repetem para aquela entidade, uma segunda fase de construção na qual irão constar atributos que podem existir várias vezes para cada entidade, e uma terceira fase em que são criados os conetores que irão relacionar os dados. Por último, uma quarta fase de construção em que é identificado os atributos dos dados.

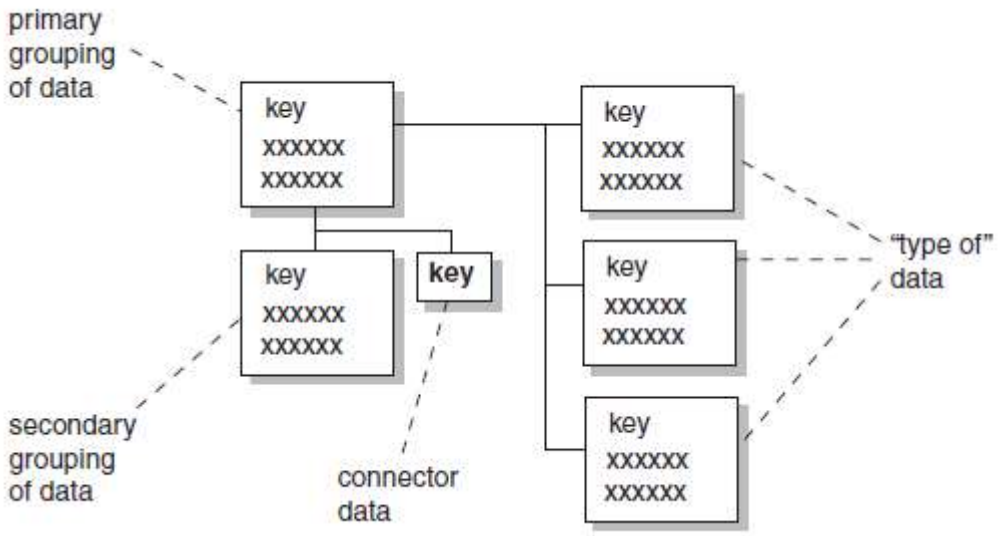

Figura 3: Construtores de segundo nível de modelação (Inmon, 2002)

Por último, um terceiro nível de modelação, a um nível mais baixo, também designado por modelo físico, que reflete o resultado do segundo nível ao qual é adicionado as chaves principais.

Este modelo requer um conhecimento elevado do negócio da empresa, assim como uma grande capacidade de entendimento e capacidade de modelagem.

1.3. RALPH KIMBALL

Ralph Kimball (Kimball & Ross, 2002) apresenta um modelo de implementação diferente da implementação tradicional de uma base de dados relacional. Usam uma metodologia designada por Modelação Dimensional, e a utilização de múltiplas Bases de Dados na sua arquitetura designadas de *Data Marts, (DM)* que são integradas

entre si através de um *Data Warehouse Bus (DWB)* de forma a garantir a coerência entre os diferentes *data marts*. Resulta assim uma metodologia que contrasta com o modelo de Imnon, principalmente na forma de construção da DW que resulta numa interpretação *Bottom-up* dos sistemas de dados da empresa.

Metodologia

Kimball (Kimball & Ross, 2002) identifica 4 passos no processo de desenho de uma data warehouse.

Selecionar a área de negocio, e o processo que se pretende modelar.

Declarar o nível de detalhe.

Escolher as dimensões que relevantes.

Identificar os factos numéricos.

O resultado será uma base de dados composta por uma tabela de factos, ligada por tabelas de dimensões. Esta carteiristas levou a que o modelo dimensional se designe também por *star schema*, uma vez que quando representadas em gráfico a sua aparência assemelha-se ao desenho de uma estrela.

Figura 4:Tabelas de facto e dimensões num modelo dimensional (Kimball & Ross, 2002)

As principais caraterísticas de uma *data warehouse* com base nesta implementação é que ela apresenta uma construção vocacionada para a consulta. As consultas tornam-se mais fáceis de elaborar, pois a implementação evita junções sucessivas entre tabelas relacionadas, e proporciona um local central onde as medidas constam em tabelas de factos que se encontram relacionadas com tabelas de dimensões que

lhes dão o contexto. Tudo isto torna o trabalho de pesquisa menos elaborado, pois exige menos conhecimentos técnicos aos utilizadores na área de pesquisa os quais normalmente têm outras aptidões em áreas de interesse diferentes de interesses informáticos.

Esta implementação, sendo muito simplista adquire uma dimensão mais complexa no momento em que vários factos são incluídos de forma a permitir diferentes data-marts que irão coexistir na mesma data warehouse. Esta evolução resulta em schemas cuja configuração vai adquirir um formato de constelação, na qual as diversas tabelas de factos relacionam-se com as tabelas dimensão que lhes dão contexto.

As tabelas de dimensões são por regra tabelas desnormalizadas, com muitos campos. São tabelas que considerando a sua relação com as tabelas de factos possuem registos que sofrem menos inserções do que os factos. No entanto existem registos que por uma questão de organização podem ser alvo de normalização, evitando assim surjam repetidas em elevado numero. Normalmente são campos cujos dados ocupam um espaço maior. Nesta situação pode ser útil normalizar a tabela de dimensão, resultando a DW numa configuração conhecida por *snow flake*.

1.4. IMPLEMENTAÇÃO TOP-DOWN VERSUS BOTTOM-UP

Paulraj Ponniah (Ponniah, 2001) descreve de forma sucinta as vantagens e desvantagens relativamente a cada uma destas abordagens: Numa adaptação livre pode-se descrever da seguinte forma:

Vantagens de uma abordagem tipo *Bottom-up*

Implementação mais fácil e rápida.

Provas de conceito e retorno de investimento mais favoráveis.

Menor risco de falha na implementação.

Incremental e escalável, o que permite a implementação de *Data Marts* após primeiras implementações.

Permite que a equipa do projeto aprenda e cresça com o mesmo.

Desvantagens:

Cada *data mart* possui a sua visão própria dos dados.

Permite data redundante nos *data mart's*

Dados inconsistentes e irreconciliáveis são perpetuados

Prolifera interfaces não gerenciáveis.

Vantagens de uma abordagem tipo *Top-Down*

Transmite uma visão corporativa dos dados.

Arquitetura unida, e não junção de diversos *data mart's*.

Armazenamento central de dados relacionados ao conteúdo.

Regras de controle centralizadas.

Se implementado com iterações, pode-se ter resultados rápidos.

Desvantagens

Demora mais tempo para ser implementado.

Maior exposição ao risco de falha na implementação.

Necessita de uma equipa de técnicos com maiores conhecimento e experiência.

Despesas mais elevadas e dificuldade de implementar ambiente de prova de conceito.

1.5. ETL - EXTRACT TRANSFORM LOAD

O processo de ETL é um dos principais pilares de uma data Warehouse, e corresponde aos processos de extrair dados das bases de dados e outras fontes relevantes o conteúdo significante, transformar através de ações de limpeza e verificação de conformidade os dados segundo os requerimentos estabelecidos e inserir esses mesmos dados numa data warehouse.

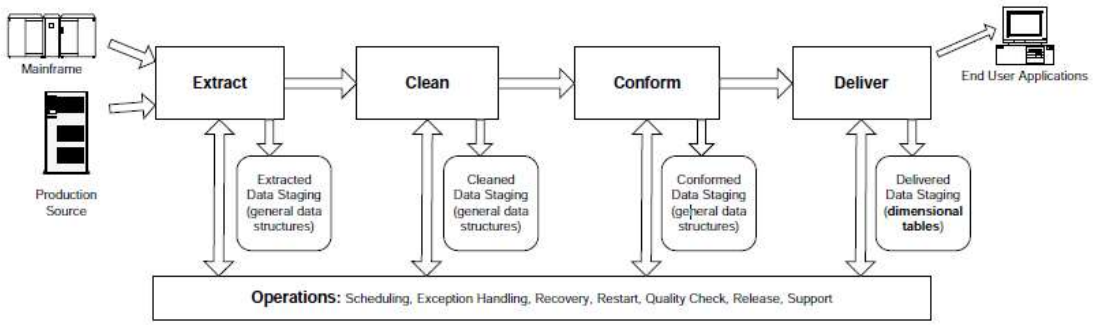

Figura 5:O processo de ETL – Fonte: (Kimball, Ralph; Caserta, Joe, 2004)

A estruturação de um processo de ETL deve começar por estabelecer os requisitos do processo. Ralph Kimball identifica os seguintes no livro Data Warehouse ETL Toolkit (Kimball, Ralph; Caserta, Joe, 2004): as necessidades do negócio; a conformidade dos dados; o perfil dos dados; a segurança dos dados; a integração dos dados; a latência dos dados; os interfaces do utilizador final; os conhecimentos técnicos; o licenciamento de sistemas; e por fim a escolha da arquitetura

Estes requisitos devem de ser definidos com clareza, de forma a que o processo de extração, transformação e carregamento de dados seja efetuado de forma transparente e os processos de ETL ficarem bem definidos.

O processo de extração deve ser incidir sobre os dados empresarias e guardados em tabelas independentes de forma serem trabalhados sem causar impactos nos sistemas operacionais.

O processo de transformação dos dados vai assim lidar com dados com origem em diferentes sistemas, cada qual com as suas próprias regras e tipos de dados que podem não obedecer ao padrão definido para a DW. Este processo incidirá assim na limpeza e transformação dos dados que não estão em conformidade com os requisitos, levando-os a um ponto de consistência. Como exemplo podemos encontrar moradas e números de telefone com diferentes formatos, códigos de produtos com diferentes anotações ou até mesmo campos numéricos com diferentes implementações. Também registo incompletos e sem interesse podem ser alvo de transformação ou até mesmo descartados se não tiverem significado relevante. Outro dos fatores importante nesta fase é a colocação dos dados na granularidade definida. Nem sempre a caraterística da data warehouse requer que os dados sejam

transpostos na granularidade com que se apresentam nas bases de dados operacionais, ou potras vezes o mesmo dado tem origem de sistemas com diferentes granularidades. Um exemplo prático é por exemplo medidas que podem estar registadas em diferentes sistemas métricos.

Segue-se o processo de carregamento dos dados em data warehouse, quando estes cumprirem os requisitos e estarem em conformidade com as necessidades.de pesquisa do utilizador final. Este passo é importante num sistema de data warehouse, porque pretende-se que sejam carregados uma única vez com o seu significado temporal, e não serem alvo de futuros atualizações e, ou correções.

1.6. STRUCTED QUERY LANGUAGE - SQL

SQL é a linguagem padrão das bases de dados, e como tal não deixa de ter um papel relevante em todo o processo relativo a data warehousing.

Para o processo de ETL recomenda-se a utilização de ferramenta desenhadas para esse propósito, uma vez que têm integradas funcionalidades que irão facilitar em muito todo o processo, tornando-o mais ágil, rápido, resultando num processo mais fácil. Este facto torna-se ainda mais relevante em sistemas de maior complexidade, e com volume de dados de diversas origens e por vezes dispares entre si na sua forma.

Não existe qualquer inconveniente numa implementação por código SQL, mas tornam-se um processo mais dispendioso e mais trabalhoso. Ainda relativamente ao SQL a extração de dados deve ser efetuada de forma a obter resultados corretos. Estão identificadas diversas lacunas que se tornam verdadeiras armadilhas no procedimento de extração de dados operacionais utilizando esta linguagem e que estão diretamente relacionados com junções entre tabelas. As mais comuns são as relacionadas com a existência de caminhos múltiplos de pesquisa sobre dados, principalmente quando se pretende a obtenção de dados agregados.

No artigo publicado no WorlsCIST 2019, 7th World Conference on Information Systems and Technologies, Cavique (Cavique, Cavique, & Gonçalves, 2019). Descreve dois principais problemas que podem ocorrer no processo de desnormalização de bases de dados relacionais com a extração de tabelas de factos. Identifica em primeiro lugar a problemática de existirem caminhos alternativo entre

tabelas de factos e tabelas de pesquisa, passando por tabelas intermédias, evidenciando assim caminhos múltiplos de acesso, que o operador *join* vai retornar valores diferentes, correspondendo às relações existentes nas tabelas intermédias. qual existem tabelas intermédias poderem retornar resultados diferentes quando se pretende retirar um facto. Considerando que uma única tabela intermédia pode não devolver todos os resultados esperados. Identifica também nas relações ternárias entre três tabelas em que existe uma relação muitos para muitos, (N:1 – 1:1 – 1:N) com recurso a uma tabela intermédia, ocorre uma inflacionamento nos dados agregados, que resulta da projeção natural do operador *join*, tornando os valores irreais. É apresentado duas formas de desnormalização de forma a evitar os caminhos múltiplos, uma primeira solução, designada por 1DF em que é proposto uma estratégia de divisão do caminho de acesso múltiplo com recurso a uma estratégia de divisão em dois caminhos distintos de pesquisa de forma a cada tabela ficar virtualmente conectada às tabelas intermédias e de pesquisa por um só caminho de acesso, evitando assim os caminhos múltiplos na pesquisa. Isto é conseguido por um processo de duplicação das tabelas onde se regista o acesso por diferentes caminhos; uma segunda técnica nomeada 2DF de desnormalização em que o objectivo é encontrar para cada facto a sua própria árvore de pesquisa. Sem caminhos duplos. E é conseguido identificando árvores independentes para cada facto identificado na base de dados, e resulta na divisão de toda a árvore do schema em diferentes árvores, de forma a que cada tabela de factos esteja relacionada com tabelas de pesquisa por caminhos de acesso independentes no que diz respeito às tabelas intermédias.

1.7. CUBOS E OLAP

Uma data warehouse é uma base de dados relacional. Apresenta, no entanto, uma subtileza funcional que a distingue das bases de dados relacionais que são projetadas para serem utilizadas nos processos de negócio e desenhadas para serem eficazes quando sujeitas a uma utilização exaustiva em processos transacionais que incluem pesquisas, inserções ou alterações em um ou vários registos, conhecidas por serem bases de dados operacionais, OLTP systems (OnLine Transaction Processing systems).

A distinção de uma data warehouse para um sistema OLTP é precisamente a sua função: fornecer informação estratégica para a tomada de decisão. Para se conseguir este objetivo os dados relevantes deixam de ser os registos atualizados como os

existentes numa base de dados operacional, mas sim dados históricos, sumários de dados e dados que se podem derivar dos que se encontram online. A estrutura de dados deixa de ser relevante para transações, e ganha importância que sejam projetadas e otimizados para consultas, que normalmente são efetuadas com uma frequência menor, mas com tempos de resposta grandes.

Uma estrutura de uma data warehouse dimensional apresenta-se como um sistema, nas quais os dados encontram-se relacionados entre valores e dimensões de conteúdos. Esta configuração apresenta um desenho lógico que permite que sejam consultados os dados com o detalhe disponível ou sumarizados, e esta facilidade de consulta torna-se uma boa fonte de alimentação para sistemas informacionais, conhecidos pela sigla OLAP do inglês OnLine Analytical Processing System cuja estrutura reflete os dados tratados de forma a poderem explorados sob diversas perspetivas. A esta nova forma de apresentar os dados que invoca a figura geométrica de um cubo designa-se por cubo OLAP, e tem a vantagem de ter os dados já organizados em células individuais, relacionadas com as dimensões relevantes para extração de informação base de suporte à decisão. Os dados num cubo OLAP podem ser explorados sob diversas perspetivas, serem fatiados, agrupados e sumarizados de forma mais rápida, e suportam de forma geral as seguintes operações:

Slice, operação que permite selecionar dados por dimensões;

Dice, operação que permite obter cubos menores através de extração de dados por dimensões;

Drill Down, operação que permite desagrupar dados de forma a se ir obtendo maior detalhe;

Roll Up, operação inversa de Drill Down.

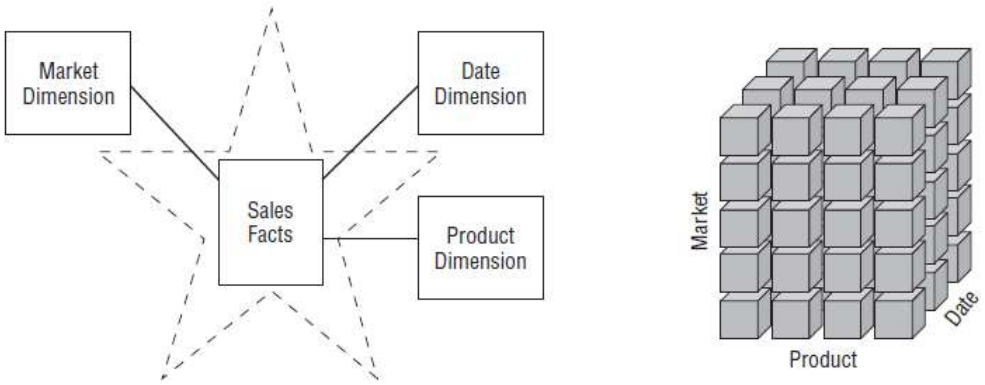

Figura 6: Star Schema versus cubo OLAP - fonte: (Kimball & Ross, 2002)

Os cubos OLAP são representações dos dados geridas por software que possuem motores próprios de indexação e agregação, estrutura essa que difere de fabricante para fabricante que concorrem entre si em inovação de forma a apresentar produtos capazes de oferecer melhores desempenhos na gestão de recursos de análise do que aqueles que são possíveis de retirar por consulta numa data warehouse que, apesar de ser vocacionada para a consulta, normalmente apresenta um volume de dados bastante elevado com tendência para crescer com o tempo, o que torna os processos de consulta, de agregação e de sumarização um processo mais demorado. Os cubos OLAP respondem com mais eficiência a estes problemas, pois podem ser estruturados com a sumarização ou detalhe pretendido para os dados factuais e relaciona-los diretamente com as dimensões relevantes para determinado tipo de questão, podendo serem oferecidos e ficarem disponíveis para exploração aos utilizadores finais de forma permanente.

Os cubos OLAP tem vindo a ganhar uma importância crescente na capacidade que têm de oferecer ao utilizador final a possibilidade de obter informação relevante para a tomada de decisão de forma mais célere e intuitiva, pois têm a possibilidades de serem construídos de forma a oferecer respostas às questões que costumam ser colocadas com maior frequência.

Capítulo 2 – Data Sources (Fontes dos dados)

Este trabalho desenvolve-se em torno de uma empresa que se dedica ao aluguer de vídeos, que será designada por VideoX.

Pretende implementar uma *Data Warehouse* que responda às necessidades da empresa de analisar o seu histórico e possuir uma ferramenta que lhe permita a tomada de decisão com bases solidas.

2.1. REFERÊNCIA PARA O ESTUDO DE CASO

Luís Cavique, et al. num artigo publicado no WorlsCIST 2020, 8th World Conference on Information Systems and Technologies (Cavique, Cavique, & Santos, 2020) propõe um processo para extração de uma data warehouse de diversas bases de dados relacionais, decompondo-as através de uma visão botton-up, objectivando a criação de um esquema de constelação (constelation schema).

O principal objetivo na extração de um data mart é identificar corretamente a tabela de facto que lhe está associada, identificar a medida que lhe está subjacente e construir uma árvore sem caminhos duplicados para as tabelas que proporcionam significado. Essa identificação é conseguida pela construção de tabelas de resumo onde se identifica o tipo de tabela e das medidas presentes, e pela elaboração de uma matriz que faz a associação entre os factos apurados e as dimensões que lhes fornecem conteúdo.

Parte deste trabalho terá como base os processos e princípios descritos nos artigos mencionados, nomeadamente na base à extração das tabelas de factos e dimensões das bases de dados.

2.2. ESTRUTURA DE DADOS

A VideoX possui na sua organização orgânica 3 departamentos, que gerem de forma independente 3 bases de dados transacionais.

O departamento cuja responsabilidade cabe o registo de alugueres de vídeo, utiliza uma réplica da base de dados sakila sample database, disponível no sitio da internet https://dev.mysql.com/doc/sakila/en/ (Oracle, 2020)

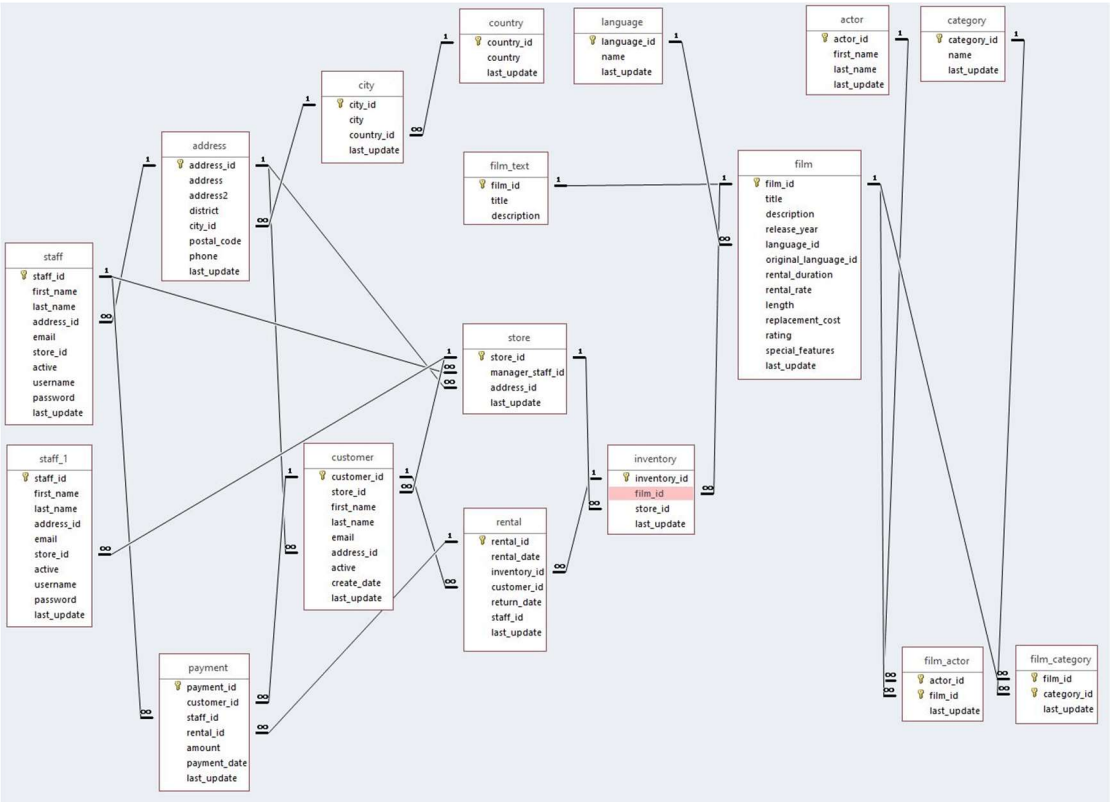

Figura 7: sakila Schema, (fonte própria)

Tabela 1: Tipo de tabelas no schema sakila

table name	type of table	type of facts
country	lookup	
language	lookup	
actor	lookup	
category	lookup	
city	intermediate	
address	intermediate	
film	intermediate	
staff	intermediate	
store	intermediate	
customer	intermediate	
inventory	intermediate	
rental	intermediate	

film_text	fact	additive
payment	fact	additive
film_actor	fact	without facts
film_category	fact	without facts

Tabela 2 : Tabela de factos do schema sakila

table name	payment	film_text	film_actor	film_category
country	X			
language	X	X	X	X
actor	X		X	
category	X			X
city	X			
address	X			
film	X	X	X	X
staff	X			
store	X			
customer	X			
inventory	X			
rental	X			
time	X	X	X	X

O departamento de recursos humanos, possui uma base de dados na qual se regista os funcionários, os departamentos a que estão afetos, os dias e os motivos de ausência bem como outras indicações pessoais, tal como contactos.

O schema da base de dados é representado pelo diagrama da Figura 8: Schema emplyees do departamento recursos humanos.

que se reproduz em anexo B com maior detalhe. Este schema é de elaboração própria, com influência no schema disponível no sítio da internet https://dev.mysql.com/doc/employee/en/.

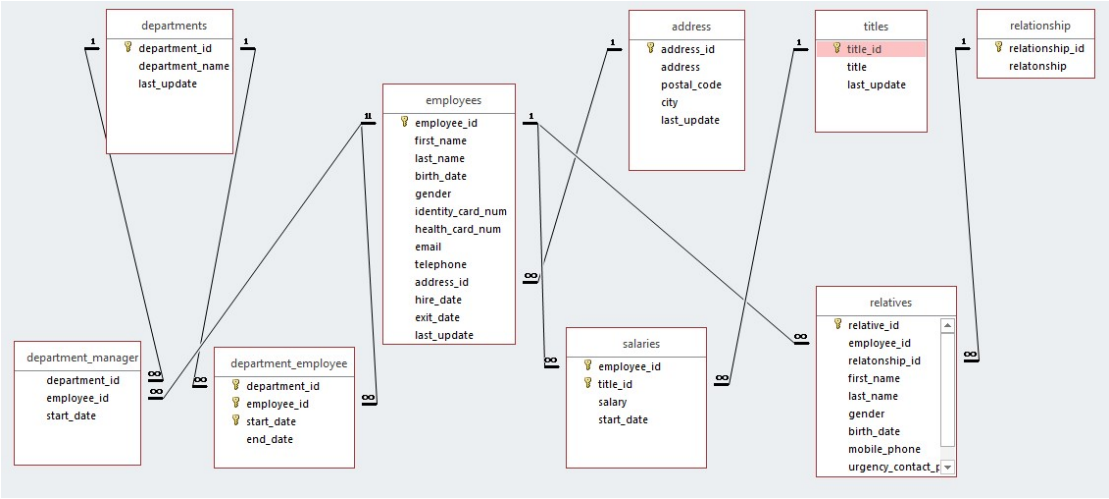

Figura 8: Schema emplyees do departamento recursos humanos.

Tabela 3: Tipo de tabelas no schema employee

table name	type of table	type of facts
departments	lookup	
address	lookup	
titles	lookup	
relatioship	lookup	
employees	intermediate	
department_manager	fact	without facts
department_employee	fact	without facts
Salaries	fact	additive
relatives	fact	without facts

Tabela 4: Tabela de factos do schema employee

table name	department_manager	department_employees	salaries	relatives
departments	X			
address	X	X	X	X
titles	X		X	
relatioship	X			X
employees	X			
time	X	X	X	X

O departamento de recursos aquisições, possui uma base de dados de elaboração própria, onde se mantém um registo dos vídeos adquiridos, referenciando o responsável pela aquisição, assim como um registo de contactos efetuados com os fornecedores.

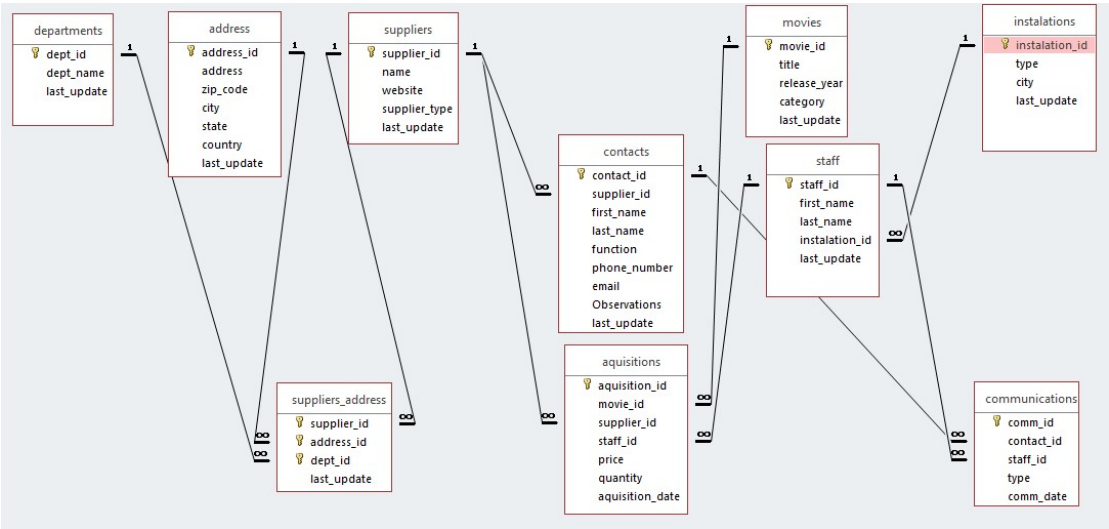

Figura 9: Schema suppliers do departamento recursos aquisições

Tabela 5: Tipo de tabelas no schema suppliers

table name	type of table	type of facts
departments	lookup	
address	lookup	
suppliers	lookup	
movies	lookup	
instalations	lookup	
contacts	intermediate	
staff	intermediate	
suppliers_address	fact	without facts
aquisitions	fact	additive
communications	fact	additive

Tabela 6: Tabela de factos do schema suppliers

table name	suppliers_address	aquisitions	communications
departments	X		
address	X		
suppliers	X	X	X
movies		X	
instalations		X	X
contacts			X
staff		X	X
time	X	X	X

Com esta informação, irá proceder-se ao estudo de caso de implementação da data warehouse da VideoX.

Capítulo 3 – Matriz de Oferta-procura da Informação

3.1 – Narrativa do SI

(a desenvolver)

3.2- Matriz Oferta-Procura

(a desenvolver)

3.3- ETL

(a desenvolver)

3.4- Perguntas OLAP (5 perguntas)

(a desenvolver)

Capítulo 4 - Conclusão

Implementação de uma data warehouse tem sido um tema amplamente debatido e estudado desde que Ralph Kimball e Bill Inmon, considerados os pioneiros, desenvolveram bases sólidas através de duas interpretações distintas de implementação, mas com o objetivo único de construir bases de dados orientadas e otimizadas para a exploração de dados. Hoje existe uma vasta gama de estudos e processos de implementação, bem como publicações que de forma mais ou menos didática abordam o assunto. Destaca-se neste campo os artigos de Luís Cavique et al. Mencionados que abordam questões práticas, em linguagem acessível e compreensível, o que vem contrastar com artigos mais técnicos e de linguagem mais difícil de compreensão para quem não está muito familiarizado com a temática.

A segunda parte deste trabalho, irá refletir-se precisamente numa implementação prática de uma data warehouse, baseado essencialmente nos artigos descritos.

Bibliografia

Adamson, C. (2010). *Star Schema, the complete reference.* McGraw-Hill.

Cavique, L., Cavique, M., & Gonçalves, A. (2019). Extraction of Fact Tables from a relacional Database: An Effort to Establish Rules in Denormalization. *7th World Conference on Information Systems and Technologies.* La Toja Island, Galiza, Spain.

Cavique, L., Cavique, M., & Santos, J. M. (2020). Supply-demand matrix: a process-oriented approach for data warehouses with constellation schemas. *8th World Conference on Information Systems and Technologies* .

Inmon, W. H. (2002). *Building the Data Warehouse, Third Edition.* John Wiley & Sons, Inc.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit second edition.* New York: John Wiley & Sons, Inc.

Kimball, Ralph; Caserta, Joe. (2004). *The Data Warehouse ETL Toolkit.* Indianapolis: Wiley Publishing, Inc.

Oracle. (1 de Setembro de 2020). *Sakila Sample Database*. Obtido de mysql: https://dev.mysql.com/doc/sakila/en/

Ponniah, P. (2001). *Data Warehousing Fundamentals: A comprehensive Guide for IT professionals.* John Wiley & Sons, Inc.

Silvers, F. (2007). *Building and Maintaning a Data Warehouse.* Taylor & Francis Group, LLC.