

Mestrado em Gestão de Informação

Master Program in Information Management

CRIAÇÃO DE UM DATA WAREHOUSE EMPRESARIAL

Na Direção de Planeamento e Gestão da TAP Portugal

André Henggeler de Moura

Relatório de Estágio apresentado como requisito parcial para
obtenção do grau de Mestre em Gestão de Informação

2017

Título: CRIAÇÃO DE UM DATA WAREHOUSE EMPRESARIAL
Subtítulo: Na direção de Planeamento e Gestão da TAP Portugal

André Henggeler de Moura

MGI



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CRIAÇÃO DE UM DATA WAREHOUSE EMPRESARIAL

por

André Henggeler de Moura

Relatório de Estágio apresentado como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Gestão do Conhecimento e Business Intelligence.

Orientador: Professor Vítor Pereira Duarte dos Santos

Novembro 2017

RESUMO

Nos dias de hoje as tecnologias de informação têm um papel fundamental no sucesso das empresas. Um dos pontos chaves para alcançar este sucesso é tentar criar uma vantagem competitiva ao analisar os dados que as próprias organizações criam.

A complexidade da informação apresentada às pessoas que geram estas identidades torna a gestão de informação um assunto chave a ter em consideração. É portanto necessário construir mecanismos ou ferramentas que agregam, processam e partilham a informação obtida permitindo uma tomada de decisão mais rápida e mais eficiente.

O principal objetivo deste estágio foi criar um Data Warehouse, para um departamento específico da empresa, com recurso a diversas fontes de informação. Um dos pontos importantes é a necessidade que as empresas têm em ajustar os próprios sistemas às bases de dados. Este tipo de aplicação permite ter uma diferente visão dos dados e oferece, também, a possibilidade da empresa ter uma análise mais forte dos dados de modo a suportar decisões.

PALAVRAS-CHAVE

Data Warehouse; Dados; Base de Dados; Business Intelligence

ABSTRACT

Information technologies have a crucial role in the success of the companies in today's world.

Among the most important weapons to reach this kind of success, creating one competitive advantage for analyzing the data that companies create in their business model is absolutely important and represents one way to do it.

The complexity of the information that is presented for who manage this kind of activities, makes the information management one issue that we all must have on count. It is necessary build either mechanisms or tools that aggregate, process and share all the information that was obtained, allowing having one decising-making more efficient and faster.

The main target of this internship was to create a Data Warehouse from several Data Sources to a specific department of the firm.

One of the most important themes is the necessity that all the companies have to adjust their owns systems to Data Bases. This sort of application allows having a different outlook of the data and on the other hand also gives the possibility to the Company to having one stronger analysis of the data in order to support it decisions.

KEYWORDS

Data Warehouse; Data; DataBases; Business Intelligence

ÍNDICE

1. Introdução	9
1.1. Contexto académico	10
1.2. Contexto empresarial	10
1.3. Objetivos do estágio	11
2. Revisão literatura.....	13
2.1. Conceito de Sistemas de Informação	15
2.2. Business Intelligence	16
2.3. Sistema de Data Warehouse	17
2.4. Processo ETL	20
2.5. Sistema de Apoio à Decisão	22
2.6. Data Mart	23
3. Tecnologias e ferramentas	27
3.1. Tecnologias.....	27
3.1.1 SQL	27
3.1.2 MDX.....	27
3.1.3 DAX.....	27
3.2. FERRAMENTAS.....	28
3.2.1 Microsoft SQL Server	28
3.2.2 Visual Studio.....	29
3.2.2.1 SSIS	30
3.2.2.2 SSAS	31
3.2.2.3 SSRS	31
4. Atividades desenvolvidas	32
4.1. Contexto	32
4.2. Projeto – Montagem de um processo ETL	33
4.2.1 Desafio	33
4.2.2 Tarefas e Skills.....	33
4.1.3 Resultados.....	36
4.2 Projeto – Criação de Relatórios.....	36
4.2.1 Desafio	36
4.2.2 Tarefas e Skills.....	36
4.2.3 Resultados.....	37

5 Conclusão.....	38
5.1 Apreciação Crítica do Trabalho Desenvolvido.....	38
5.2 Apreciação Global do Estágio e Aprendizagem.....	38
5.3 Perspectivas Futuras.....	39
Referências Bibliográficas	40

ÍNDICE DE FIGURAS

Figura 1 – Organograma da TAP Portugal em Junho de 2016	11
Figura 2 - Relações entre dados, <i>capta</i> , informação e conhecimento (Adaptado de Checkland e Holwell 1998).....	15
Figura 3 – Exemplo do processo ETL	21
Figura 4 – Exemplo de um Star Schema.....	24
Figura 5 – Exemplo de um Snowflake Schema.....	25
Figura 6 - Estrutura típica e simplificada de um sistema tecnológico de BI.	26
Figura 7 – Ambiente de desenvolvimento do SQL Server.....	28
Figura 8 – Ambiente inicial do Visual Studio.....	29
Figura 9 – Ambiente de desenvolvimento do SSIS.....	30
Figura 10 – Exemplo de um cubo multidimensional.....	31
Figura 11 – Modelo tipo de um processo ETL.....	33
Figura 12 – Processo de extração da informação para uma tabela fonte.	34
Figura 13 - Processo ETL de carregamento de dados para uma tabela de staging.	35

LISTA DE SIGLAS E ABREVIATURAS

BD	Base de Dados
BI	Business Intelligence
DAX	Data Analysis Expressions
DM	Data Mart
DPG	Direção de Planeamento e Gestão
DW	Data Warehouse
ETL	Extract, Transform, Load
MDX	Multidimensional Expressions
OLAP	Online Analytical Processing
SA	Staging Area
SAD	Sistema de apoio à Decisão
SI	Sistemas de Informação
SSAS	SQL Server Analysis Server
SSIS	SQL Server Integration Server
SSRS	SQL Server Reporting Server
SQL	Structured Query Language
TAP	Transportes Aéreos Portugueses
TI	Tecnologias de Informação

1. INTRODUÇÃO

O presente relatório descreve as tarefas que desempenhei durante a realização do Estágio Curricular do Mestrado de Gestão de Informação, com especialização em Gestão do Conhecimento e Business Intelligence, da Universidade NOVA - IMS.

O relatório encontra-se organizado em cinco partes. Na primeira parte apresenta-se um enquadramento do estágio e da empresa onde desenvolvi o trabalho apresentado, a TAP Portugal. Na segunda parte dá-se um enquadramento teórico às tarefas por mim desenvolvidas e uma abordagem a alguns conceitos essenciais usados ao longo do trabalho. No terceiro grupo encontram-se as tecnologias utilizadas ao longo do estágio. Em seguida procura-se descrever as tarefas realizadas, sem esquecer o produto final do estágio, a criação do Data Warehouse (DW) empresarial. Por fim é feita uma breve análise crítica ao estágio, às tarefas que desempenhei e às dificuldades com que me deparei ao longo do tempo.

O conceito de uma Base de Dados (BD) está associado a um enorme volume de dados. Nos dias de hoje a velocidade com que a TAP gera dados é incalculável e é cada vez mais importante o seu armazenamento para otimizar qualquer tipo de atividade. A maior dificuldade passa por tentar armazená-los de uma forma organizada e de fácil consulta. O tempo acabou por se tornar um verdadeiro aliado dos sistemas de informação ao mostrar que uma base de dados é a melhor e mais eficaz forma de guardar dados, independentemente do seu volume.

1.1. CONTEXTO ACADÉMICO

O estágio relatado teve como foco a organização e a gestão de dados dentro da TAP Portugal. O presente relatório de estágio visa destacar as atividades realizadas no âmbito do Mestrado de Gestão de Informação, com especialização em Gestão do Conhecimento e *Business Intelligence* (BI) pela Universidade Nova de Lisboa – IMS – *Information Management School*. O estágio teve início a 09 de Dezembro de 2015 e foi concluído no dia 03 de Junho de 2016, havendo decorrido nas instalações da TAP Portugal, no Aeroporto de Lisboa na Direção de Planeamento e Gestão (DPG).

Antes da era tecnológica o processo de armazenamento de dados da TAP era doloroso e pouco eficaz com desvantagens evidentes para os utilizadores e não possibilitava acesso rápido à informação. Atualmente, o uso do computador é uma prática totalmente enraizada dentro da empresa. A necessidade de guardar dados nos computadores deve-se ao crescimento exponencial da informação que é preciso armazenar ou de que convém ter registo.

Neste contexto, a utilização de uma Base de Dados veio solucionar diversos problemas dentro na TAP.

1.2. CONTEXTO EMPRESARIAL

A TAP Air Portugal conhecida como TAP Portugal é a maior companhia aérea portuguesa e uma das maiores da Europa, fundada em 1945 com sede no Aeroporto de Lisboa. É, desde Março de 2005, membro integrante da *Star Alliance*, a primeira e a maior aliança de companhias aéreas do mundo.

A TAP assume a missão ter como foco o serviço de Transporte Aéreo e atividades afins, aspirante a ser sempre a melhor opção para quem utiliza os seus serviços e uma das melhores empresas para se trabalhar. Para tal – à data de 1 Junho de 2016 – opera em média 2500 voos por semana e dispõe de uma frota moderna com mais de 77 aviões.

Desde o primeiro dia e durante todo o período de estágio na TAP, foi-me atribuída, como principal missão, a entrega de informação com elevada qualidade e fidedigna para posterior análise com o objetivo de otimizar processos e reduzir custos.

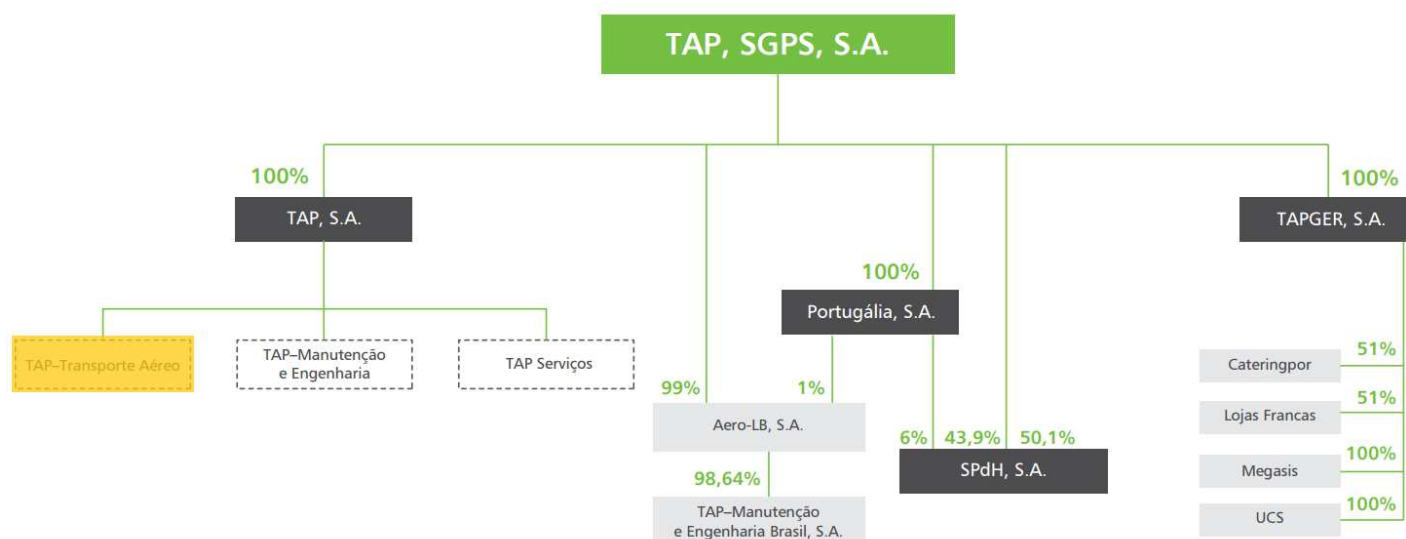


Figura 1 – Organograma da TAP Portugal em Junho de 2016

Como estagiário foi-me dada a oportunidade de integrar a área de operações de voo, Transporte Aéreo – ver figura 1 - mas sempre com a total liberdade de contactar com outros sectores e projetos. No entanto, o presente relatório trata apenas do projeto onde estive envolvido desde o início e o qual considero se destaca entre os demais.

1.3. OBJETIVOS DO ESTÁGIO

O principal objetivo do estágio foi criar um Data Warehouse para a área de operações de voo. Para alcançar o objetivo final teve de se passar por algumas etapas.

Em primeiro lugar foi necessário entender a importância da arquitetura de uma base de dados. A forma como as tabelas estão dispostas e interligadas tem uma grande importância

e é uma mais-valia quando se procura aumentar a eficácia na extração de um grande volume de informação.

Em seguida, de forma a simplificar a análise dos dados, procedeu-se à montagem do processo de extração, transformação e carregamento de dados, em inglês *Extract-Transform-Load* – ETL, que permite reunir informação de diversas fontes de dados numa única base de dados. Este processo possibilita que os utilizadores ganhem tempo nas suas ações e tenham a informação toda concentrada num único local.

De seguida escreveram-se algumas *querys* em linguagem *Structured Query Language* – SQL, para que os utilizadores possam obter a informação pretendida de forma rápida e eficaz não perdendo assim tempo na construção das mesmas.

Para que os utilizadores finais possam aceder à informação pretendida têm apenas de aceder ao Data Warehouse criado, ao invés de terem que procurar a informação em diversas fontes de informação e diversos programas o que se traduz numa perda tempo e eficácia, podendo, muitas vezes, ter informações contraditórias e repetidas. Com a criação deste DW o departamento ganhou alguma maturidade e obteve resultados visíveis num espaço de temporal mais curto e com menos investimento. No entanto, é de realçar que a constante atualização dos processos de ETL que alimentam o DW em questão não deve ser uma opção a colocar de parte, uma vez que proporciona uma visão da organização como um todo, em detrimento da visão fragmentada que existia.

2. REVISÃO LITERATURA

Atualmente o mundo está em constante mudança. Mais do que nunca as inovações surgem a uma velocidade muito grande e todos os dias aparecem novos produtos e novas ideias. O próprio planeta Terra é rico em mudança. No mundo dos negócios, claramente, a mudança é também palavra de ordem: qualquer organização que não se adapte às transformações do seu meio envolvente será ultrapassada facilmente. De forma a acompanhar um mercado empresarial cada vez mais competitivo como o atual é muito importante que as empresas evitem ao máximo todos os possíveis problemas que possam conduzir ao insucesso nas suas iniciativas. Com o resultado da evolução das Tecnologias de Informação (TI), bem como a capacidade de armazenamento dos computadores, hoje em dia praticamente todas as empresas utilizam sistemas informáticos para suportar os seus projetos de negócio.

A definição de *Bentley e Whitten* (2007), apresentada na obra *Systems Analysis & Design Methods*, define TI como:

“Uma combinação de pessoas, dados e processos que interagem para recolher, processar, armazenar e fornecer como resultado a informação necessário para apoiar uma organização”

Com o passar do tempo, estes sistemas acabam por gerar uma enorme quantidade de dados relacionados com o negócio (Porter e Millar, 1985). Estes dados que estão armazenados, são um recurso que, de forma geral, não são utilizados pelas organizações (Subramanian, Smith e Nelson, 1996). Efetivamente as bases de dados não são projetadas para produzir informações estratégicas, o que torna os sistemas inapropriados para o apoio à tomada de decisão (Gupta, 1997).

Uma base de dados é o “local” onde se podem guardar vários registos, dados, com o propósito de armazenar a informação de uma forma estruturada e organizada onde permite que o utilizador procure essa mesma informação de uma forma rápida e eficaz.

O conceito de uma BD está associado a um enorme volume de dados. Nos dias de hoje a velocidade com que se geram dados é incalculável e a grande dificuldade passa por tentar armazená-los de uma forma organizada. O auge do processo atinge-se quando se consegue

transformar uma quantidade enorme de dados em informação e posteriormente em conhecimento.

O tempo acabou por se tornar um verdadeiro aliado dos sistemas de informação ao mostrar que uma base de dados é a melhor e mais eficaz maneira de guardar a informação, independentemente do seu volume (Brobst, 1998).

Citando uma das frases mais conhecidas de Charles Darwin (1809-1882) relativamente à importância da adaptação:

“Na evolução das espécies não são os mais fortes que sobrevivem mas sim o que mais rapidamente se adaptam às mudanças”.

2.1. CONCEITO DE SISTEMAS DE INFORMAÇÃO

Existem vários conceitos que contribuem para a definição de Sistema de Informação (SI). Checkland e Holwell (2014) distinguem quatro conceitos base: dados, *capta*, informação e conhecimento. Estes conceitos podem ser definidos da seguinte forma:

- Dados: Factos não estruturados;
- Capta: Dados relevantes a que escolhemos prestar atenção;
- Informação: Atribuição de um significado aos dados já selecionados;
- Conhecimento: Vai além das informações pois além de ter um significado tem uma aplicação

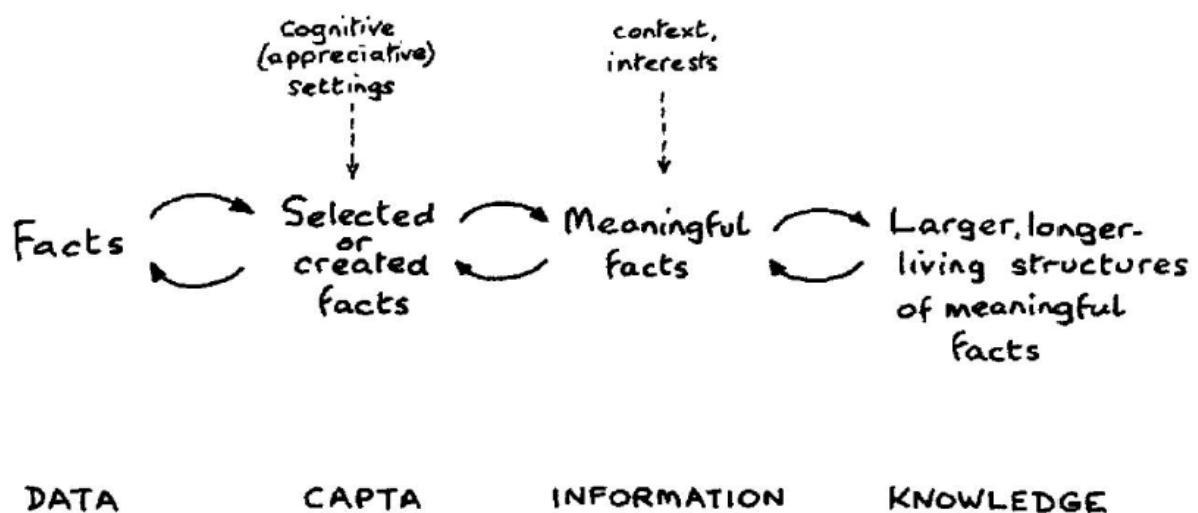


Figura 2 - Relações entre dados, *capta*, informação e conhecimento (Adaptado de Checkland e Holwell 1998)

Com os conceitos bases já presentes é possível agora definir o que se entende por SI, “*pode ser definido como um conjunto de componentes interrelacionados que trabalham em conjunto para agrupar, recuperar, processar, armazenar e distribuir informação para facilitar o planeamento, controle, coordenação, análise e processo decisório nas organizações.*” (Laudon e Laudon 2014).

Os computadores passaram a ser fundamentais na melhoria das atividades de processamento de informação organizacional desde que começaram a ser utilizados para o processamento de dados. Os SI baseados em computadores passaram de um mero suporte periférico para um fator chave de uso comum ao longo da totalidade da organização (Boland e Hirschheim, 1987).

2.2. BUSINESS INTELLIGENCE

Os sistemas de *Business Intelligence* são, hoje em dia, um catalisador da mudança permitindo uma boa gestão, ajudam na tomada de decisão, tornam a empresa mais ágil num mercado cada vez mais competitivo, onde a recolha e análise de dados é a chave do sucesso. Um sistema de BI engloba um vasto conjunto de aplicações de apoio a tomada de decisão o que possibilita um acesso rápido e partilhado das informações bem como a sua análise.

O termo *Business Intelligence* foi utilizado pela primeira vez por Luhn (1958) como a “*a habilidade de apreender as inter-relações dos factos apresentados, numa forma que permita guiar a ação rumo a um objetivo desejado*”

Em 1989 o termo é popularizado por Howard Dresner como denominador geral para descrever um conjunto de conceitos e métodos para melhorar a tomada de decisão do negócio através da utilização de sistemas de suporte à decisão (Power, 2007).

Para Wu (2007) o termo BI é utilizado para descrever um conjunto de aplicações e tecnologias que são utilizadas para recolher, aceder e analisar dados de uma organização com o objetivo de ajudar a tomar decisões melhores. A finalidade de *business intelligence* é fornecer informação utilizável.

O âmbito de BI cobre processos relacionados com tecnologia bem como processos relacionados com o negócio (Schwenk, 2010) tendo English (2005) proposto as seguintes definições:

- **BI:** "Capacidade de uma empresa para agir eficazmente através da exploração dos seus recursos humanos e de informação.";

- **Ambiente de BI:** Informação de qualidade contida em bases de dados bem desenhadas que fornecem aos utilizadores acesso atempado e de forma intuitiva às informações, permitindo tomar decisões corretas.

Business Intelligence permite aos membros de toda a organização não só analisar informações de negócio com o objetivo de melhorar a qualidade de informação mas também a melhoria de processos operacionais. (Elbashir, 2008)

2.3. SISTEMA DE DATA WAREHOUSE

Um Data Warehouse não é nenhum tipo de aplicação que se possa comprar e instalar nos computadores da empresa. Na realidade, a sua implementação exige a integração de vários processos. Numa perspetiva minimalista, um DW não é mais que uma BD especializada, integra que gere a recolhe de informação a partir de sistemas transacionais internos e fontes de dados externas. Um DW é um repositório de dados separados das bases de dados operacionais da organização, definido como “uma coleção de dados orientados por assunto, integrados, variáveis no tempo e não-voláteis, que suporta o processo da tomada de decisão” (Inmon, 2002)

Um DW é contruído para permitir: uma vista integrada e completa de toda a organização; acessos a dados históricos da organização; ter uma fonte de dados verosímeis dentro da organização e facilitar os processos de tomada de decisão, sem sobrecarregar os sistemas operacionais (Por, Klauer e Brobst, 1998). Todavia, terá sido Bill Inmon – conhecido como “o pai do data warehousing” – que popularizou o conceito de Data Warehouse, e que providenciou os critérios de definição hoje comumente aceites (Inmon, 2005).

Segundo *Bill Inmon* (2002) o DW é definido como:

“... a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management’s decisions.”

Um Data Warehouse é um repositório de dados que compreende as seguintes características:

- i. **Orientado por assunto** - todas as entidades e transações relacionadas com um determinado assunto - por exemplo, vendas, recursos humanos ou hospitais - estão ligadas entre si;
- ii. **Integrado** - Inmon refere que este é, de facto, o aspeto mais importante de um data warehouse (Inmon, 2005); os dados que alimentarão o DW são oriundos de vários sistemas fonte, e por isso necessitam de ser limpos, filtrados e transformados, para que possam representar uma visão integrada do negócio;
- iii. **Não-volátil** - os dados que são carregados para o data warehouse, por norma, não sofrem alterações ou atualizações. Ao contrário dos sistemas operacionais, apenas durante as operações de carregamento (periódicas planeadas) é que os dados são alvo de algum tipo de edição;
- iv. **Variável no tempo** - ao contrário de um sistema transacional, em que os registos apenas estão associados ao momento de acesso (são válidos 'agora'), um data warehouse mantém registos que estão associados a todos os momentos onde ocorreu uma alteração nos dados (dados históricos).

Um DW bem contruído tem um impacto muito significativo a vários níveis ao permitir descobrir novas formas de aumentar a receita e a eficiência da organização, bem como aumentar a qualidade do suporte à decisão.

Existem duas diferentes abordagens mais debatidas, *top down* e *bottom up*, para o desenho de uma DW que foram introduzidas por Inmon (2002) e Kimball (2002), respetivamente. A "*Corporate Information Factory*" de Inmon (2002), corresponde à abordagem *top down* que defende a existência de um repositório central com um modelo de dados normalizados, para alimentar os Data Marts dependentes com vista a permitir a exploração de dados numa lógica departamental. Já Kimball (2002) tem outra perspetiva, *bottom up*, onde consiste na criação de Data Marts que podem ser combinados entre si para formar a DW da organização através de um interface de ligação ("*bus*"), tempo Kimball chamado a esta arquitetura "*Data Warehouse Bus Architecture*".

No entanto existem outras abordagens que podem ser utilizadas onde se adotam princípios das duas abordagens (Burstein 2008).

A necessidade de obter informação estratégica, a partir de um grande volume de dados dispersos levou que fosse desenvolvido um novo género de Sistema de Informação (SI) designado de Data Warehouse, estes SIs são contruídos com o intuito de apoiar o processo de tomada de decisão na organização (Boar, 1997). O conceito de DW remonta ao final dos anos 80 onde era tido com sendo “o único repositório lógico de toda a informação de negócio utilizada para efeito de reporting” (Devlin & Murphy, 1988).

A definição de *Ralph Kimball* (2002), apresentada na obra *The Data Warehouse Toolkit*, define DW como:

"... a copy of transaction data specifically structured for query and analysis."

Com o surgimento dos DWs foi necessário criar novos métodos de estruturação de dados, tanto no armazenamento como na consulta de informação. As empresas produzem e armazenado um volume elevado de dados, sendo normal que estes dados estejam dispersos por várias servidores, que podem, inclusive, estar dispersos por várias localizações geográficas e ter sido desenvolvidos em plataformas e linguagem diferentes (Gupta, 1997).

Um dos principais desafios da implementação de um DW é a integração dos dados, eliminando redundâncias, identificando os duplicados que possam estar em sistemas distintos, representadas sob formatos ou designações diferentes (Adelman e Moss, 2000).

O processo de passagem dos dados dos sistemas transacionais para o DW é denominado *Extraction, Transformation and Loading*. O sistema de DW, normalmente, é separado das bases de dados transacionais, pelo que as consultas dos utilizadores não degradam a performance dos sistemas transacionais, que ficam simultaneamente protegidos de alterações e perdas causadas pela manipulação indevida da informação.

Há vários anos que organizações perceberam a riqueza dos dados existentes nos seus sistemas sejam eles uma folha de Excel, bases de dados dispersas, entre outros, e começaram cada vez mais a investir em sistemas analíticos. Antes de se analisar todos estes conjuntos de sistemas, nomeadamente em Business Intelligence, tem de ser ter um conjunto de dados coerente com toda a informação da empresa, o Data Warehouse.

A construção de um Data Warehouse é cada vez mais uma realidade e uma opção empresarial importante, uma vez que se trata de uma ação que envolve toda a organização e por outro lado requer um fluxo de investimento grande, quer em meios materiais, quer em meios humanos.

Uma vez contruído o DW é possível fragmenta-lo em BD mais pequenas, orientadas por assuntos ou, neste caso, departamentos, que são reconhecidos como Data Marts.

Um Data Mart (DM) é uma estrutura de dados multidimensional orientada por assunto ou por departamento que normalmente tem enfoque numa área de marketing, comercial, financeira, entre outras (Kimball et al., 1998). O seu tempo de conceção é mais reduzido. A opção de contruir um Data Mart esta relacionada com as necessidades dos utilizadores finais ou das necessidade especificas do negócios. Neste caso em particular os dados existem, a informação existe mas é necessário juntar tudo numa única base de dados. Este processo de aglomerar a informação toda irá possibilitar que todos os trabalhadores da área em questão tenham acesso à informação de uma forma mais ágil e rápida (Malinowski e Zimányi, 2009).

2.4. PROCESSO ETL

A sigla ETL vem do Inglês, *Extract Transform Load*, ou seja, Extração Transformação e Carga. Uma solução ETL permite realizar operações que passam pela extração a partir das fontes, transformação e inserção dos dados na DW e um motor OLAP - *Online Analytical Processing* – para a realização de consultar, ferramentas de análise e outras aplicações que gerem o processo de recolha e entrega de dados de forma útil e relevante aos utilizados (Cyrn, 2003). As ferramentas OLAP permitem aos utilizadores efetuarem consulta de dados através de aplicações. Ajuda a entender as diversas necessidades de negócio que os dados capturam, levando a um processo de decisão baseado em factos (Burstein 2008). Permite capturar os dados através de medidas de dimensões formando o chamado cubo OLAP (Valacich e Scheneider 2011).

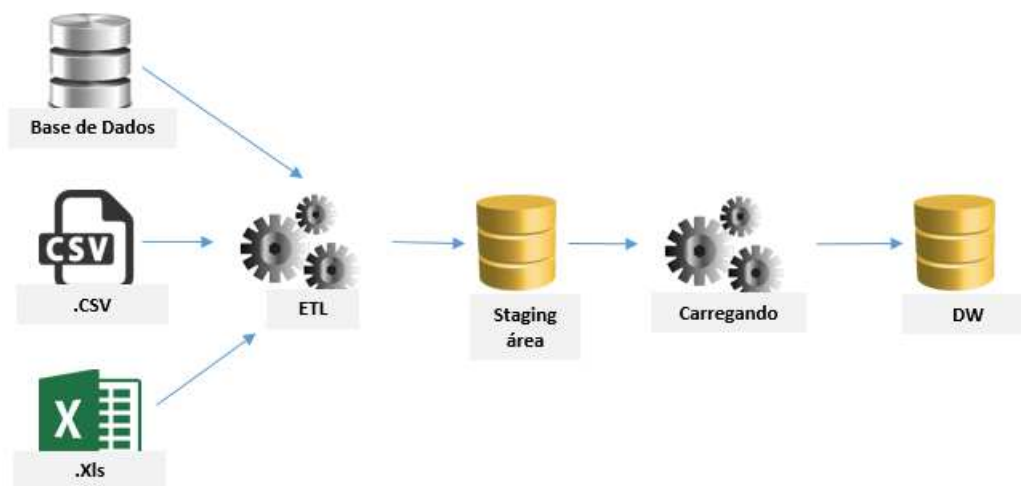


Figura 3 – Exemplo do processo ETL

O ETL é o processo mais crítico e demorado na construção de um DW que consiste em três etapas:

- 1. Extração (Extract):** É a coleta de todos os dados dos sistemas de origem que podem ser as mais diversas fontes de dados como por exemplo o Excel, um *flat file* ou outra base de dados. Depois de os extrair o sistema vai transferi-los para a *Staging Area* onde o sistema ETL pode operar independentemente das fontes de dados;
- 2. Transformação (Transformation):** É neste etapa que se realizam os devidos ajustes para tentar melhorar a qualidade dos dados extraídos. Vem a parte do tratamento da informação e da limpeza de dados, que consiste em padronizar os dados em tamanho e tipo, substituição de caracteres estranhos, correção de erros de digitação, evitar duplicidade de informações, substituição de dados não preenchidos, deve-se também tratar as diferentes definições de informação, por exemplo, ter um sistema de definição de H para homem e M para mulher e outro sistema tem M de masculino e F de feminino não é correto. Em casos como estes deve-se optar pela definição mais usual e torna-la padrão para evitar inconsistências nos dados;
- 3. Carga (Load):** A terceira e última etapa consiste em estruturar e carregar os dados para a DW. Dependendo das necessidades da organização este processo pode variar

amplamente. Algumas organizações optam por substituir as informações existentes pelos dados acumulativos e outras optam por adicionar os dados em falta no intervalo de tempo a definir. O intervalo de tempo definido por cada empresa depende das necessidades de negócio.

Na figura 3 pode-se observar um exemplo de um processo ETL onde são os dados são extraídos de três fontes diferentes, um base de dados, um ficheiro Excel e um CVS também conhecido por *Flat File*. Depois da extração os dados são enviados para a *Staging Area* (SA) onde são limpos e trabalhados. A *Staging Area* consiste numa área de armazenamento intermediária, entre as fontes de dados e o próprio DW (Kimball e Ross 2002). Posteriormente é feito o carregamento dos dados na DW final.

Se os dados em questão não forem trabalhados de uma maneira correta no processo de extração as informações geradas através dos mesmos farão com que as decisões tomadas estejam erradas, o que pode afetar diretamente a organização. Portanto, os dados devem representar a verdade (Kimball, 1998). A maior parte do esforço exigido ao longo da criação de uma base de dados é consumido no processo ETL (Inmon 1997).

2.5. SISTEMA DE APOIO À DECISÃO

Os Sistemas de apoio à Decisão (SAD) fazem parte de um conjunto de Sistemas de Informações que regem empresas em todo o mundo. Os SAD diferem dos SI porque estes redirecionam o planeamento estratégico dentro de uma empresa.

Um Sistema de Apoio à Decisão cobre uma grande variedade de sistemas e tecnologias. Estes sistemas foram criados com o intuito de servir de auxílio às pessoas que tomam decisões dentro de uma organização. Os SAD são usados em decisões, em que um julgamento é necessário, ou em decisões de que não são completamente suportadas por algoritmos, ou seja, é sempre preciso um utilizador para usar este tipo de sistemas, onde este representa o papel mais importante.

É importante realçar outras definições de um SAD:

"Os Sistemas de Apoio à Decisão têm quatro características principais: incorporam simultaneamente dados e modelos; foram concebidos para assistir os gerentes no seu processo de decisão sobre tarefas semiestruturadas; suportam julgamentos ao nível da gerência; o seu objetivo é melhorar a efetividade das decisões, e não a eficiência com que cada decisão está a ser feita." (Turban 1990)

"Os sistemas de apoio à decisão são sistemas de informação interativos que assistem o gestor a aceder a problemas mal estruturados oferecendo modelos analíticos e acesso a bases de dados." (Zwass 1992)

Um SAD trás bastantes vantagens como: ultrapassar limites cognitivos (através do computador), decisões mais eficazes e eficientes, separa as operações de produção das operações de decisão e rapidez ao longo de todo o processo.

2.6. DATA MART

Um Data Mart é um subconjunto de um DW orientados para os requisitos específicos de um departamento ou em função de um negócio, razão pela qual é muitas vezes construído tendo em consideração as necessidades dos utilizadores finais. Qualquer departamento poderá ter o seu DM, sendo que todos os dados que constam no DM são proveniente do DW, se bem que em menor escala e com uma quantidade significativa agregados (Inmon 2008).

As três principais características dos Data Marts são segundo Oppel (2004):

- Foco num departamento;
- Normalmente não contém dados operacionais;
- Implementação típica com recurso ao *Star Schema*

Os DMs são contruídos com base em modelos de dados multidimensionais, em que os dados são representados em forma de cubo e onde, normalmente, seguem um dos seguintes esquemas (Cyran 2003, Kimball & Ross 2002, Oppel 2004):

- *Star Schema* (Esquema em estrela): Desenvolvido por Ralph Kimball, consiste numa tabela de factos com várias tabelas dimensionais associadas em forma de estrela;

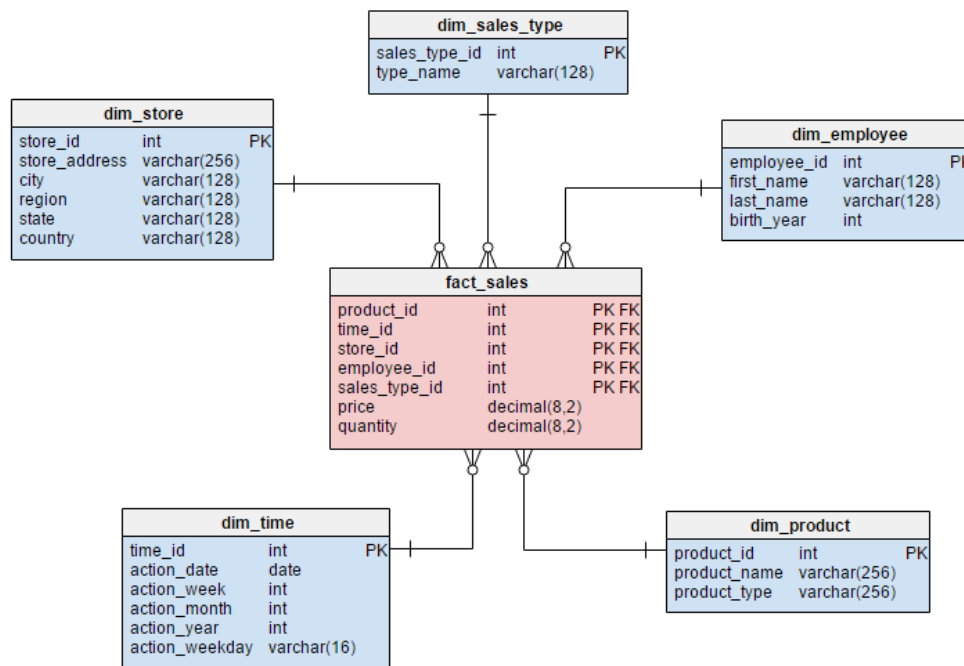


Figura 4 – Exemplo de um Star Schema.

- **Snowflake Schema** (Esquema Floco de Neve): Neste esquema as dimensões podem ter sub-dimensões próprias. Ao fazer uma normalização das dimensões no esquema em estrela obtém-se o esquema em floco de neve;

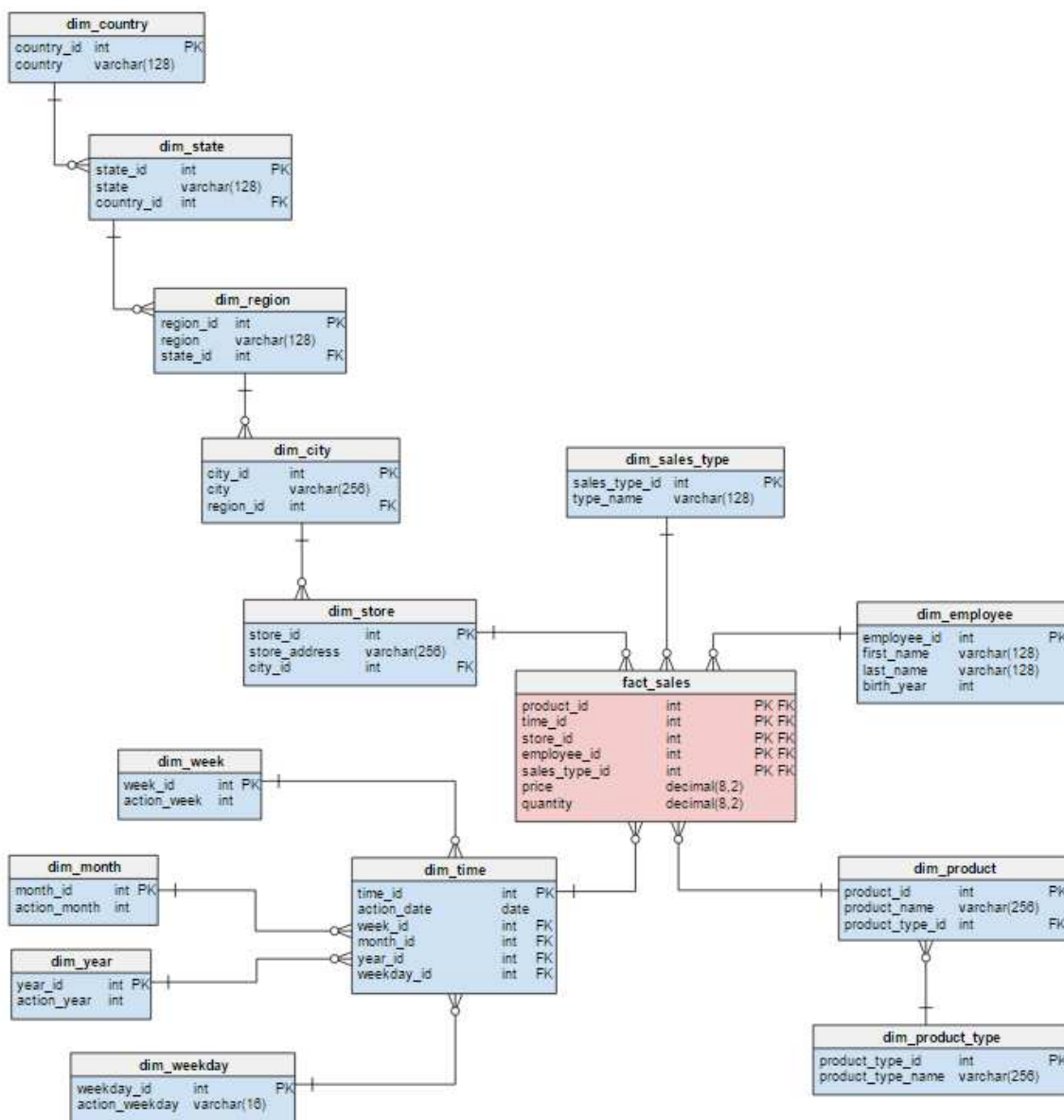


Figura 5 – Exemplo de um Snowflake Schema.

- **Constellation** (Constelação): É um conjunto de tabelas de factos que partilham tabelas dimensionais, podendo, no entanto causar limitações às potencialidades de *queries* do DW.

Uma tabela factual contém medidas e métricas, como por exemplo a quantidade vendida de um determinado produto, que correspondem aos factos do negócio. As tabelas dimensionais contém atributos que descrevem o negócio, por exemplo o cliente em questão, contextualizando os valores presentes na tabela de factos (Kimball e Ross, 2002).

Agora que se analisou a arquitetura de todo o sistema até chegarmos ao *Data Marts* consegue-se facilmente entender que o sistema de BI não está restrito a uma tecnologia específica e, por isso, a sua ideia, normalmente, torna-se mais confusa do que realmente é. O sistema de BI contém uma certa abstração no seu conceito, permitindo, assim, flexibilidade e adaptações a cada novo projeto. A estrutura final do sistema vai depender do contexto em que a solução está inserida.

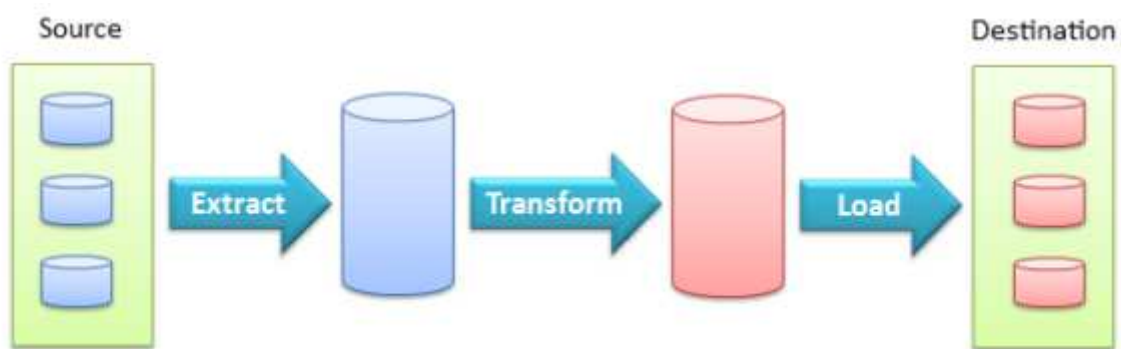


Figura 6 - Estrutura típica e simplificada de um sistema tecnológico de BI.

Na figura 6 ilustra-se ver uma estrutura simplificada de um sistema de BI desde a escolha das fontes de informação a utilizar, o processo ETL, a agregação de toda a informação no DW, a divisão dessa informação em *Data Marts* e a chegada do conteúdo pretendido ao utilizador final.

3. TECNOLOGIAS E FERRAMENTAS

3.1. TECNOLOGIAS

3.1.1 SQL

Structured Query Language (SQL) é uma linguagem de programação em que o principal objetivo é a manipulação, controlo, transação e consultas de dados. Deste modo tem como função ser a interface entre o utilizador e o sistema gestor da base de dados seja, por exemplo, SQL Server ou Oracle. Esta linguagem é utilizada para modelos relacionais.

3.1.2 MDX

Multidimensional Expressions (MDX) é uma linguagem bastante complexa, algo semelhante ao SQL, mas destinada a modelos multidimensionais. O MDX tem a capacidade de manipular e analisar a informação em *online analytical processing cube*. Os cálculos de métricas, KPI's e construção de relatórios com recurso ao SQL Server Reporting Services - SSRS – requerem a utilização desta elaborada linguagem.

3.1.3 DAX

Data Analysis Expressions (DAX) é uma nova linguagem de fórmulas que permite que os utilizadores definam cálculos personalizados. O DAX inclui algumas das funções utilizadas em fórmulas Excel e funções concebidas para trabalhar com dados relacionais e efetuar a agregação dinâmica. De uma forma geral esta linguagem ajuda o utilizador a gerar informações novas a partir de dados já existentes nos modelos.

3.2. FERRAMENTAS

3.2.1 MICROSOFT SQL SERVER

O SQL Server é um sistema que gere bases de dados relacionais, desenvolvido pela Microsoft. A sua principal função é de armazenar dados que sejam fornecidos por outros *softwares*. A figura 7 mostra o ambiente de desenvolvimento do SQL Server.

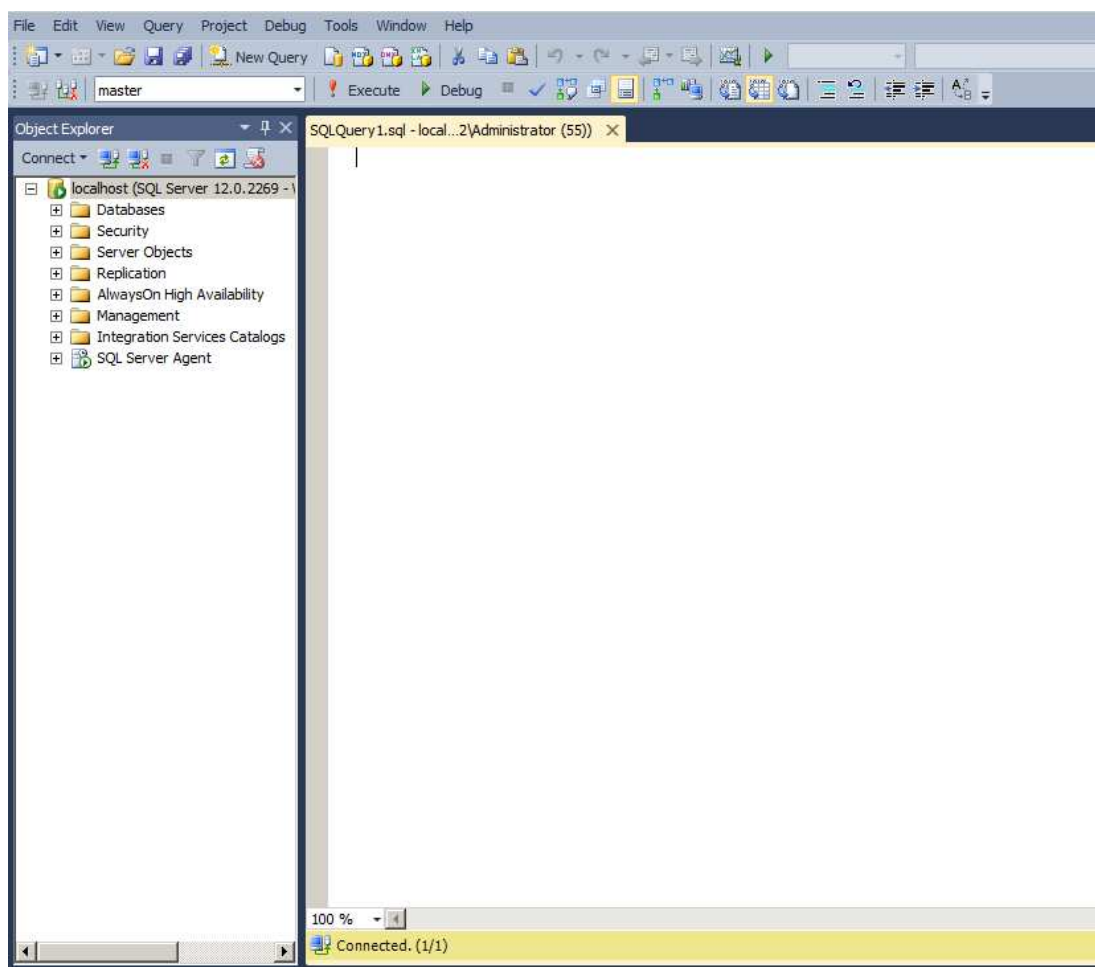


Figura 7 – Ambiente de desenvolvimento do SQL Server.

É possível a visualização, manipulação e análise de dados através desta ferramenta com o auxílio de *queries*. O SQL Server encontra-se dividido em quatro módulos: Database Engine, Integration Services - SSIS, Analysis Services - SSAS e Reporting Services - SSRS.

3.2.2 VISUAL STUDIO

O Visual Studio é um conjunto de programas criado pela Microsoft. Esta ferramenta é conhecida como uma *Integrated Development Environment*, um software com um editor de texto muito poderoso.

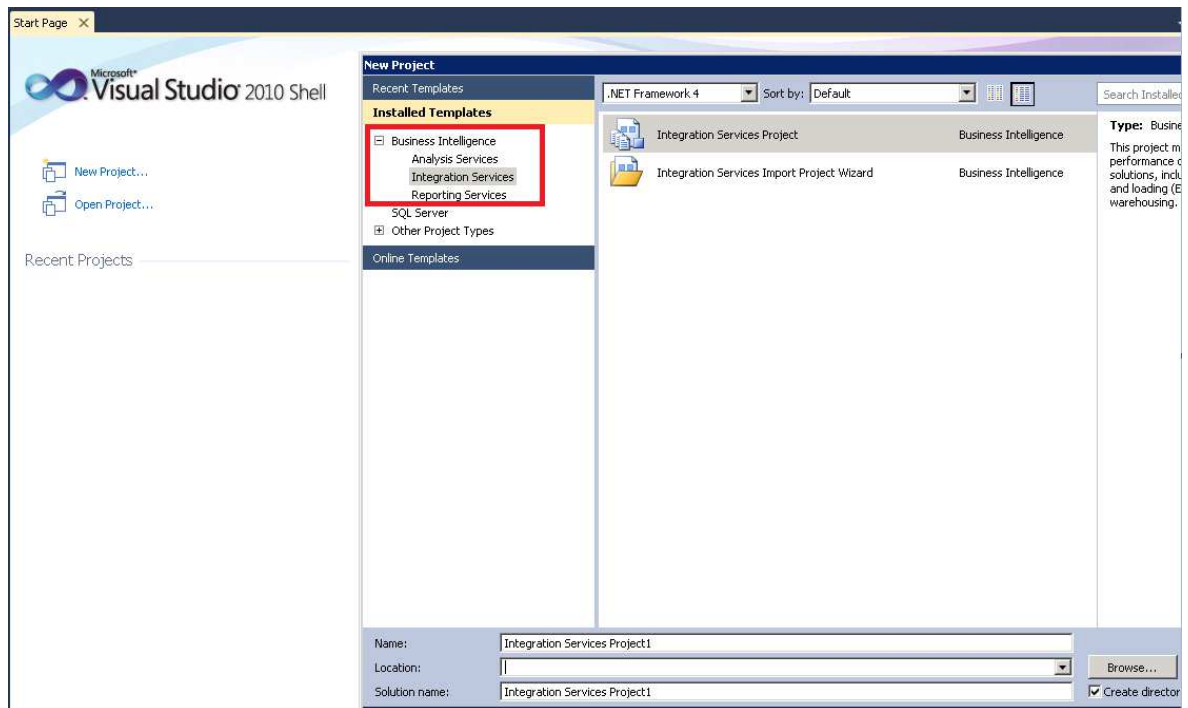


Figura 8 – Ambiente inicial do Visual Studio.

Na figura 8 pode-se observar o seu ambiente. Além da componente de programação integrada também os três módulos, SSIS, SSAS e SSRS, já referidos do SQL Server.

3.2.2.1 SSIS

SQL Server Integration Service (SSIS) permite a integração de dados com alta *performance* e construção de ETL, constituindo várias operações para a construção de um Data Warehouse. O SSIS inclui algumas ferramentas gráficas para a criação de tarefas pré-definidas como por exemplo, execuções de instrução SQL, ligação às fontes de informação, envio automático de emails, transformações e manipulação de dados, agregações, entre outras.

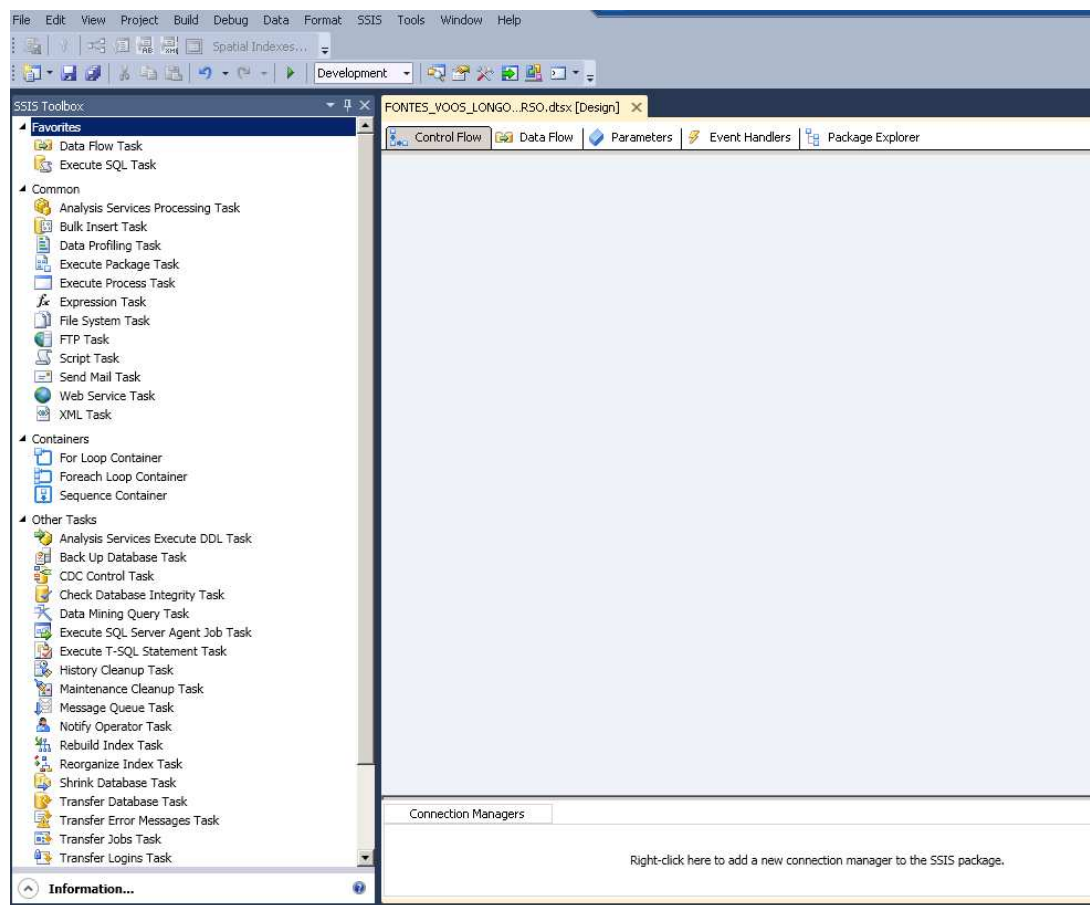


Figura 9 – Ambiente de desenvolvimento do SSIS

Na figura 9 podemos ver, do lado esquerdo, algumas tarefas existentes do SSIS, estas tarefas apesar de pré-definidas necessitam de uma configuração exaustiva de acordo com o tipo de dados, respetivas fontes de informação e objetivo final do utilizador.

3.2.2.2 SSAS

SQL Server Analysis Server (SSAS) possibilita a construção de on-line Analytical Processing Component (OLAP) do SQL Server. O SSAS permite construir estruturas multidimensionais, os chamados cubos, e armazenar agregações complexas. O *Analysis Services* é uma visão multidimensional da nossa base de dados.

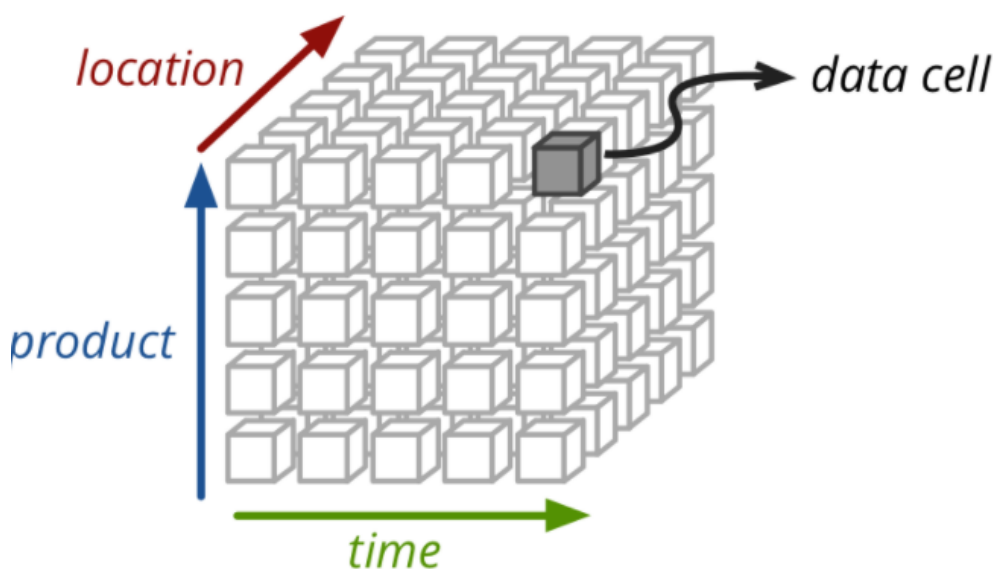


Figura 10 – Exemplo de um cubo multidimensional.

3.2.2.3 SSRS

SQL Reporting Service é um módulo do SQL responsável pela construção de relatórios que tem por base os dados provenientes do cubo já construído. Em todo este módulo a linguagem usada é o MDX. O SSRS é bastante utilizado nos dias de hoje visto que comporta ações desde construção de tabelas, a relatórios que são expansíveis e permitem exibir categorias e subcategorias de produtos até apresentar dados de uma forma gráfica.

4. ATIVIDADES DESENVOLVIDAS

4.1. CONTEXTO

O projeto apresentado em seguida foi desenvolvido na sua totalidade, em todas as fases, por mim. No decorrer deste estágio foram realizadas as atividades abaixo descritas.

O projeto consistiu em criar de raiz um Data Warehouse tendo por base várias fontes de informação. Até ao momento todos os dados utilizados pela DPG eram consultados acedendo a uma serie de programas e, posteriormente, compilados de forma manual, num único Excel. Só depois deste processo demorado é que era feita uma análise a toda a informação, que muitas vezes se revelava pouco precisa e muito demorada.

Tendo em conta este cenário o maior pedido feito pela equipa da TAP prendia-se com o facto de ter a informação disponível de forma imediata e com o maior rigor possível.

O estágio foi dividido em dois grandes projetos. O primeiro projeto passou pela criação de um processo ETL, com o intuito de construir o DW. No segundo, desenvolveram-se alguns relatórios, tendo por base o DM já existente, para que a informação possa estar sempre disponível e atualizada.

Sem grande conhecimento da área de negócio, mas com uma grande vontade de aprender, tive inicialmente algumas dificuldades para perceber que tipo de informação era fundamental e de que forma o departamento queria que os dados fossem apresentados. Está é uma das principais dificuldades encontradas por alguém que não tem o *know-how* da área.

4.2. PROJETO – MONTAGEM DE UM PROCESSO ETL

4.2.1 DESAFIO

O projeto tinha um tempo limitado para a sua realização o que constituiu um enorme desafio visto que o volume de dados gerados por dia na TAP Portugal é astronómico. As várias bases de dados englobam muita informação e os dados necessitam de ser analisados antes de efetuar a sua extração. Para a implementação desta solução foram necessárias, na fase inicial, bastantes reuniões com toda a equipa para elaborar não só o desenho de todo o processo como perceber de que forma era trabalhada a informação a que tinham acesso. Um dos grandes desafios foi tentar otimizar o mais possível todos os processos para que o tempo de espera dos utilizadores fosse o menor possível. Para tal foram sempre realizados vários testes e montagem de vários processos em paralelo com o objetivo de chegar a uma solução ótima.

4.2.2 TAREFAS E SKILLS

A informação passava muito pelo tipo de aviões que fazem determinadas rotas, as horas de voo de cada membro da tripulação e as horas voadas por cada avião, são exemplos de informação que era recorrentemente pedida. Todas estas informações estavam em repositórios de dados variados, desde ficheiros em Excel, bases de dados SQL e Oracle e ficheiros CSV. O primeiro passo foi perceber onde é que a informação se encontrava alocada e de que forma se conseguiam ligar os dados das várias fontes sem alterar o seu contexto. De seguida construiu-se o processo ETL que passou por três fases: fontes, *staging* e analítico. Na figura 11 pode observar-se um modelo típico de um processo de ETL.

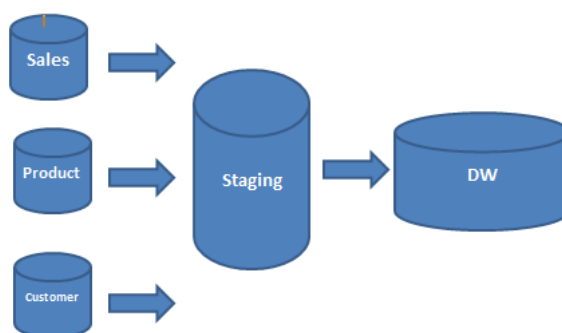


Figura 11 – Modelo tipo de um processo ETL.

A passagem da informação das fontes externas para as tabelas designadas por fontes são uma cópia direta dos dados. Neste passo apenas queremos garantir que os dados já se encontram do nosso lado para que os possamos trabalhar no próximo passo. Sempre que existe um carregamento dos dados para as tabelas da DPG os dados são todos apagados e voltados a carregar. No processo de fontes não é guardado nenhum tipo de histórico. É um processo rápido com o objetivo apenas de garantir apenas a cópia dos dados de um lado para o outro. Na figura 12 temos um exemplo concreto do que foi descrito. Como se pode observar no final do carregamento temos duas tarefas, uma de sucesso e outra de falha. Para existir um controlo maior dos dados que são carregados é inserido numa tabela de *logging* as horas que o processo termina e o nome do *package* do SSIS em questão. Mais tarde caso seja necessário fazer algum controlo ou se algum dos processos der erro é mais fácil identificar em que passo foi, para que se possa corrigir o eventual erro de uma forma mais simples e eficaz.

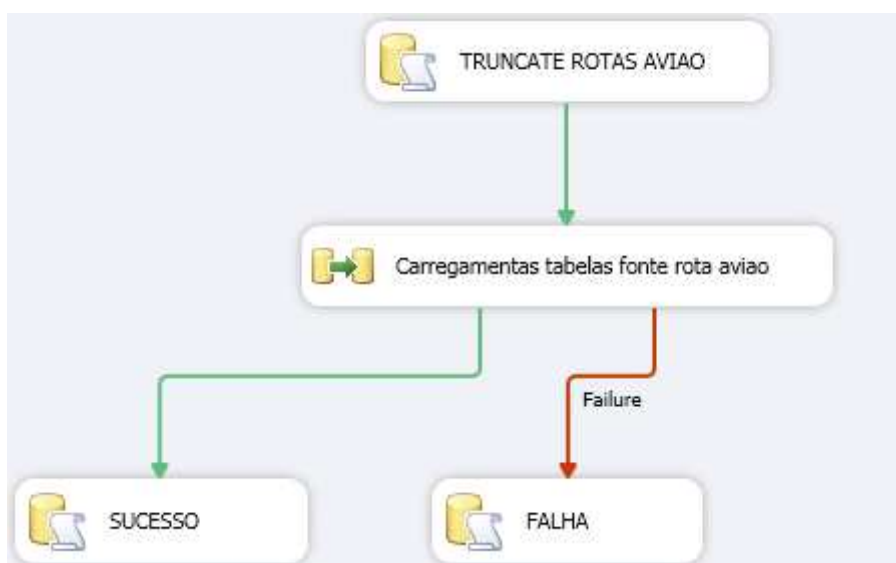


Figura 12 – Processo de extração da informação para uma tabela fonte.

Assim, após termos os dados pretendidos todos guardados nas nossas tabelas fontes passou-se para as tabelas de *staging*.

É neste passo que se faz o tratamento da informação consoante as regras de negócio que foram pedidas pela DPG.

Podemos observar na tabela 13 um exemplo da passagem da informação, onde é feito um tratamento mais exaustivo dos dados. Nesta passagem é guardado um historio de toda a informação que recebemos. Por motivos de confidenciais não irá ser explicado em detalhe o processo apresentado mas de uma forma geral é feita uma divisão entre os tipos de aviões e os respetivos destinos. Por fim a informação é colocada numa tabela com o nome “STG_TP_RATting_AIRBUS”

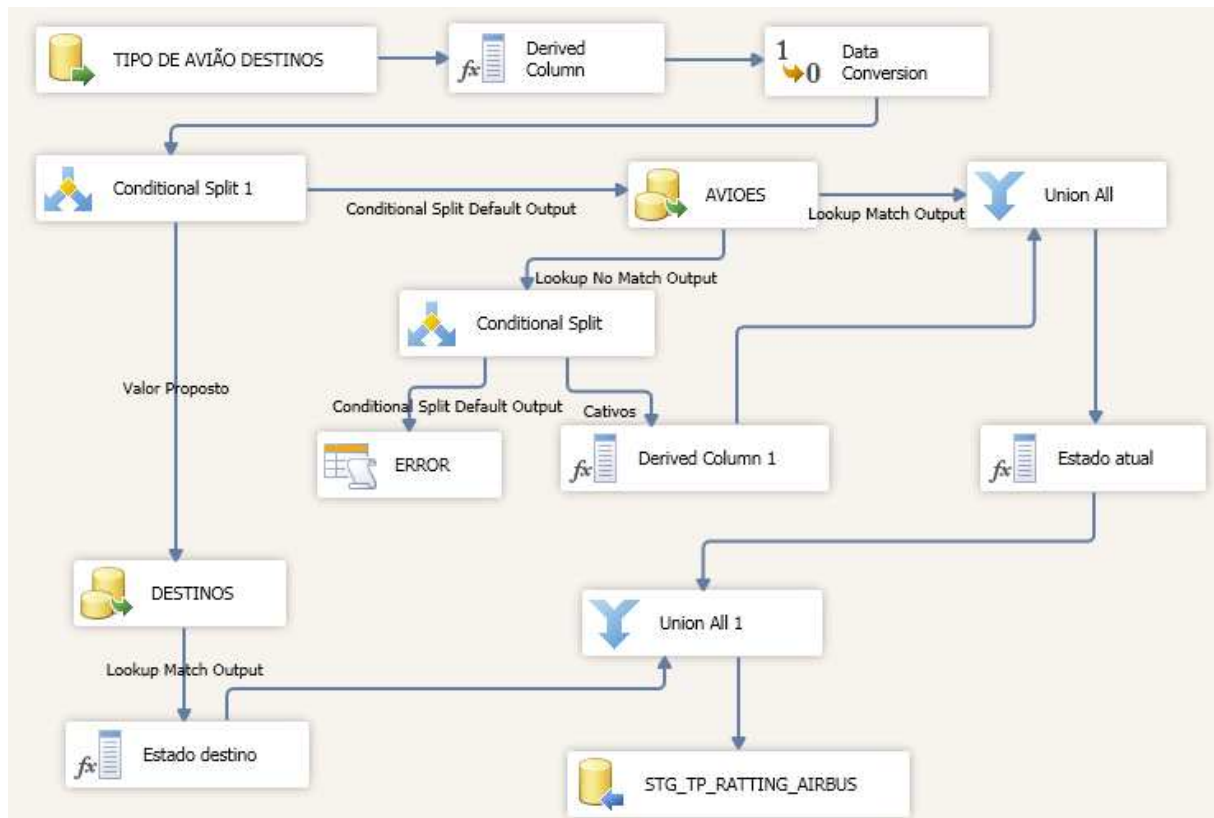


Figura 13 - Processo ETL de carregamento de dados para uma tabela de *staging*.

Após termos a informação totalmente trabalhada e guardada nas tabelas de *staging* estamos em condições de passarmos para o último passo.

Na última etapa fazemos uma cópia da informação que consta na *staging area* para o DW final.

Todos os processos que foram montados estão inseridos em vários *jobs* para que corram de forma automática e os utilizadores tenham a informação atualizada diariamente.

4.1.3 RESULTADOS

A elaboração e a implementação de uma solução de Business Intelligence, tendo por base a construção do referido processo ETL revelou ser uma mais-valia, tanto para a organização como para os utilizadores, na optimização da diminuição dos erros gerados ao longo de todo o processo, como também no tempo que era perdido pelos colaboradores a montar as recorrentemente análises.

As ferramentas de BI demonstraram que são uteis e uma mais valida para qualquer negócio e/ou empresa que possua uma grande estrutura de dados para a qual o tempo de análise e estruturação dos dados seja limitada.

4.2 PROJETO – CRIAÇÃO DE RELATÓRIOS

4.2.1 DESAFIO

Após a realização do projeto anterior faltava realizar a última etapa; a criação de relatórios. A ferramenta utilizada foi o Visual Studio SSRS e as linguagens de programação foram o SQL e o MDX.

O desafio passou pela criação de relatórios com informação sempre atualizada, tendo por base alguns relatórios já criados em Excel. Até ao momento todos os relatórios eram criados de forma manual e atualizados diariamente e alguns várias vezes ao dia. Sem um conhecimento profundo na área mas já com algum *know-how* sobre o negócio tive de entender quais as métricas mais indicadas a utilizar.

Para a elaboração dos relatórios havia a necessidade de aplicar alguns filtros e o conteúdo dos mesmos era feito maioritariamente através de matrizes ou tabelas. Foi-me dado um relatório modelos, onde já estavam definidas as cores, o logotipo e o tipo de letra a utilizar.

4.2.2 TAREFAS E SKILLS

Na maioria dos relatórios efetuados utilizou-se a informação já organizada e hierarquizada dos cubos, linguagem MDX, mas em alguns relatórios foi necessário recorrer à linguagem de SQL por uma questão de *performance*.

Para a construção de qualquer relatório é necessário a existência de *Data Sources* de onde são extraídos os dados com a informação requerida. Todas as ligações foram efetuadas ao cubo e à DW. Na criação dos *Data Sets* foram executadas *querys* para extrair os dados pretendidos, tanto para as tabelas a apresentar nos relatórios, como para o auxílio de alguns parâmetros.

Ao longo de execução deste projeto participei em diversas reuniões para entender quais as necessidades de todos os utilizadores de forma a tentar reunir todos os requisitos pedidos. Depois de os relatórios estarem finalizados foram realizados vários testes para garantir que tudo estar em condições e sem erros de forma a serem colocados em produção e prontos a serem utilizados.

4.2.3 RESULTADOS

A criação de relatórios utilizando as técnicas de Business Intelligence revelara-me uma vez mais superiores à já ultrapassada criação de documentos Excel. Na ótica de todos os utilizadores este novo método de *reporting* veio acelerar o processo de consulta de informação e diminuir consideravelmente a taxa de erros visto que o tratamento da informação é feito de forma automática.

5 CONCLUSÃO

5.1 APRECIÇÃO CRÍTICA DO TRABALHO DESENVOLVIDO

Durante a elaboração do primeiro projeto, construção do processo de ETL, foi-me possível consolidar toda a formação académica e aprofundar os meus conhecimentos na ferramenta destinada à integração de dados. Fiquei substancialmente mais familiarizado com todas as *tasks* disponíveis no programa e aprendi quando e como é que as terei de utilizar, o que acabou por me permitir uma maior eficácia na execução do projeto.

No segundo projeto trabalhei diretamente sob informação já disponível numa base de dados da empresa. Tive a oportunidade de evoluir significativamente na construção de relatórios utilizando a tecnologia SSRS. O meu conhecimento de MDX era muito limitado, o que exigiu de mim um estudo e um esforço adicional. Este projeto exigiu mais da minha parte, na medida, que só eu estava encarregue pela execução dos relatórios, ganhei autonomia total sobre o trabalho que estava a desenvolver.

Estes desafios ajudaram-me a encarar a realidade dos projetos com tempos limitados para a sua execução e a gerir todos os inconvenientes de trabalhar sob pressão. Apesar de tudo tive o privilégio de ter uma equipa fantástica ao meu lado que se mostrou sempre presente quando foi necessário. No final tive o privilégio de ver o meu trabalho elogiado por toda a equipa e utilizadores, sentindo assim, que os meus objetivos foram cumpridos.

5.2 APRECIÇÃO GLOBAL DO ESTÁGIO E APRENDIZAGEM

Desde o primeiro dia que iniciei as minhas funções, quanto estagiário, na TAP que fui acompanhado pelos melhores especialistas na área de BI. Ao longo de todo o processo fui tendo vários desafios, com necessidades de aprendizagens diferentes, o que colocou à prova a minha capacidade de resolução de problemas e organização. É fundamental referir que um dos maiores desafios, e o mais complicado de gerir, foi o facto de estar a estagiar oito horas por dia e a frequentar o mestrado ao mesmo tempo. É algo muito exigente, no entanto, considero uma experiência muito enriquecedora que me permitiu obter ensinamentos

valiosos para o resto da vida, obrigando-me a fazer uma gestão do meu tempo de uma forma minuciosa e até a descobrir capacidades que até então desconhecia.

No término desta etapa, sinto que todo o meu esforço ao longo desta jornada foi recompensado, por consegui alcançar todos os objetivos propostos, tanto por mim, como aqueles que me foram exigidos ao longo do tempo. Em suma, estou muito motivado e preparados para ultrapassar os próximos desafios que surgiram na minha profissional e para progredir na área de Business Intelligence.

5.3 PERSPECTIVAS FUTURAS

O começo desta etapa representou uma grande mudança na minha vida, não só por ter deixado a minha área de formação, mas também por ter abraçado algo completamente novo para mim, o que me cativou desde o primeiro dia que iniciei o mestrado.

Apensar dos meus conhecimentos há um ano sobre a matéria de BI serem escassos, após a conclusão deste estágio enquanto profissional acredito ter adquirido o *know-how* suficiente para continuar a fazer parte de grandes projetos como os que tive a possibilidade de participar.

Esta experiência deu-me uma enorme vontade de continuar a aprender, crescer e a expandir o meu conhecimento para conseguir aplicar as melhores praticas na vida profissional e ter uma experiencia que me permita autonomia total, passando a exercer a minha profissão como um especialista de Business Intelligence.

No meu ponto de vista a área de BI tem um grande potencial de crescimento e existe uma procura cada vez maior no mercado, o que me leva a crer acreditar que a importância da área irá cada vez ser maior nas próximas décadas.

REFERÊNCIAS BIBLIOGRÁFICAS

Boar B. H. (1997). *"Building, Using, and Managing the Data Warehouse-Understanding Data Warehousing Strategically"*, Prentice Hall PTR

Gupta, V. R. (1997). *"An Introduction to Data Warehousing"*, Ed., 1997

Inmon, W.H. (1994). *"Using the Data Warehouse"*, John Wiley & Sons, Inc.

Inmon, W.H. (1997). *"Managing the Data Warehouse"*, John Wiley & Sons, Inc.

Inmon (2005), *Building the Data Warehouse*, Wiley Publishing, 4th, Indianapolis.

Kimball R., Reeves L., Ross M E Thornthwaite W. (1998). *"The Data Warehouse Lifecycle Toolkit - Expert Methods for Designing, Developing, and Deploying Data Warehouses"*. John Wiley & Sons, Inc.

Kimball, R., & Ross, M. (2002). *The data warehouse toolkit : the complete guide to dimensional modeling*. New York: Wiley Computer Publishing.

Kimball R., Ross M., Thornthwaite W., Mundy J. e Becker B. (2008). *"The Data Warehouse Lifecycle Toolkit, 2nd ed."*. Indianapolis: Wiley Publishing, Inc.

Kimball, Reeves, Ross & Thornthwaite (1998), *The Data Warehouse Lifecycle Toolkit Expert Methods for Designing*

Malinowski & Zimányi (2009), *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*, Springer Berlin Heidelberg, Berlin.

Marjanovic (2007), *The Next Stage of Operational Business Intelligence: Creating New Challenges for Business Process Management*. IN IEEE (Ed.) *Proceedings of the 40th Hawaii International Conference on System Sciences (2007)*. Hawaii.

Moss & Atre (2003), *Business Intelligence Roadmap - The Complete Project Lifecycle for Decision-Support Applications*, Addison-Wesley.

Porter, M. E. e Millar, V. E. (1985). "How Information Gives You Competitive Advantage", Harward Business Review

Subramanian A., Smith L. D. e Nelson A. C. (1996). "*Strategic Planning for Data Warehousing in the Public Sector*", 29th Annual Hawaii International Conference on System Sciences, Hawaii, 54-60