INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Business Intelligence em Avaliações de Alunos da Licenciatura em Engenharia Informática

GISELA ESMERALDA AGUIAR DO COUTO
Outubro de 2016

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Business Intelligence em Avaliações de Alunos da Licenciatura em Engenharia Informática

Gisela Esmeralda Aguiar do Couto

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Sistemas Computacionais

Orientador: Paulo Jorge Machado Oliveira

Co-orientador: Ângelo Manuel Rego e Silva Martins

Júri:

Presidente:

[Nome do Presidente, Categoria, Escola]

Vogais:

[Nome do Vogal1, Categoria, Escola]

[Nome do Vogal2, Categoria, Escola] (até 4 vogais)

Porto, Outubro 2016

Resumo

Numa instituição de ensino, tal como em qualquer organização, é importante monitorizar o desempenho dos diferentes processos de negócio, especialmente com o objetivo da melhoria contínua. Numa instituição de ensino há processos de negócio "comuns", mas os processos de aprendizagem revestem-se da maior importância. É fundamental analisar o desempenho destes processos, estudando os comportamentos específicos de unidades curriculares e/ou subgrupos de alunos e agir em conformidade para contribuir na melhoria da aprendizagem.

Por outro lado, a monitorização também é importante para conseguir comprovar o bom funcionamento dos processos perante entidades avaliadoras e de acreditação, como a Ordem dos Engenheiros, ENAEE ou A3ES. Estas entidades solicitam um conjunto de informação estatística sobre variados aspetos da atividade letiva, com o objetivo de analisar o desempenho e o bom funcionamento, classificando-a e, se for caso disso, reconhecê-la com prémios de excelência.

Atualmente, o Diretor da Licenciatura de Engenharia Informática monitoriza a informação curricular dos seus alunos a partir de ficheiros em formato de folha de cálculo, extraídos do Portal da Instituição. Dada a complexidade de análise e de propensão a erros que este formato pode induzir, foi encontrada a necessidade de criação de um sistema de *Business Intelligence* que permita o carregamento, armazenamento e manutenção de dados, de forma mais simples e automatizada.

Depois de definidos então os objetivos para a criação de um sistema de armazenamento, o próximo passo passou por fazer um levantamento do estado de arte, definindo alguns conceitos considerados mais importantes para compreensão de toda a dissertação, estudando arquiteturas possíveis e algumas das ferramenta que foram utilizadas para o desenvolvimento da solução (ferramentas ETL e ferramentas possíveis de utilizar para futura apresentação de dados OLAP), concluindo-se com uma análise atual do mercado, apresentando algumas soluções criadas para o contexto escolar.

No contexto de avaliações de alunos, é apresentada uma possível solução desenhada para a resolução do problema, sendo esta baseada num sistema com a capacidade de armazenamento de grandes volumes de dados e respetiva manutenção de histórico. Foram feitos testes para comprovar a credibilidade do sistema, não só pelo carregamento das fontes de dados disponibilizadas, mas também pela criação de um conjunto de análises identificadas como as mais utilizadas pelo Diretor da Licenciatura.

Palavras-chave: Business Intelligence, Armazém de dados, ETL, OLAP

Abstract

In the context of a school system, like any organization, it's important to monitor the performance of the different business processes, especially to support and improve steadily. The school system has common business processes, but the most important are learning processes. It's crucial make an analisys about the processes performance, studying school specific behaviors and acting accordingly to improve learning experience.

On the other hand, monitoring is also important to be able to check the correct functioning of the school system towards evaluative and accreditation institutions, such as the "Ordem dos Engenheiros", ENAEE or A3ES. These entities require a set of statistical information on various aspects of teaching activity, in order to analyse the performance and operation, classifying and, where appropriate, recognise it with excellence awards.

Currently, the Director of Informatics Engineering Degree monitors its student's curriculum information through files in a spreadsheet format, extracted from the institution website. Given the complexity of analysis and error-prone that this format can induce, was found the need to create a Business Intelligence system that allows loading, storage and maintenance of data, more simple and in an automated way.

After all solution goals were defined to create the data warehouse, the next step passed to survey the state of the art, defining some concepts considered more important to understanding the whole dissertation, studying possible architectures and some of the tool that have been used for the development of the solution (ETL tools and possible tools to use for future presentation data OLAP), concluding with a current market analysis, showing some solutions created for the school context.

In the context of student's evaluations, a possible solution designed to solve the problem is presented below, which is based on a system with the ability to store large amounts of data and related history maintenance. At last, tests to prove the credibility of the system have been made, not only by the loading of the available data sources, but also by creating a set of identified analyses as the most used by the Director of the Degree.

Keywords: Business Intelligence, Data Warehouse, ETL, OLAP

Agradecimentos

Manifesto o meu apreço a todos aqueles que, de alguma forma, contribuíram nestes três anos para a minha formação, para a conclusão desta dissertação e sobretudo os que me ajudaram a crescer e a ultrapassar os obstáculos que foram aparecendo ao longo desta longa caminhada.

Aos meus orientadores, Professor Doutor Ângelo Manuel Rego e Silva Martins e Professor Doutor Paulo Jorge Machado Oliveira, pela orientação que deram permitindo que me fosse guiando pelo caminho certo, por toda a disponibilidade com que me ajudaram, pelas opiniões e críticas sobre o projeto, pelo incentivo em continuar até ao fim e por toda a aprendizagem e conhecimentos partilhados. Sem esta base, não seria possível a realização deste trabalho.

Não posso deixar de agradecer também ao Diretor do Mestrado Doutor Nuno Alexandre Pinto da Silva e aos restantes docentes do Departamento de Engenharia Informática que lecionaram os módulos da unidade curricular aos quais frequentei pelo incentivo, pela pronta ajuda no esclarecimento de dúvidas e partilha de conhecimento.

Aos meus amigos e colegas de mestrado por toda a preocupação demonstrada, pela motivação que recebi em continuar em frente e por todos os sorrisos partilhados nos momentos mais difíceis. Em especial agradeço ao meu colega e amigo Tiago Pereira pelo trabalho conjunto desenvolvido, pelos conhecimentos partilhados, pelo incentivo, força e sobretudo pela amizade.

Aos meus pais e irmãos o meu profundo agradecimento por me darem a oportunidade de chegar até aqui, pelo apoio perante os desafios, força nos momentos menos bons, pela alegria e sobretudo pela compreensão sempre que estive ausente por todo este trabalho. Obrigada por terem feito de mim aquilo que sou hoje.

Ao meu namorado pelo apoio ao longo deste percurso, pela confiança e valorização depositada no meu trabalho, pela ajuda e pela coragem para ultrapassar cada obstáculo. Obrigada pelo carinho incondicional, pela amizade, pela compreensão quando estive ausente, mas sobretudo obrigada por todos os momentos felizes que me proporcionaste.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

BI	*Business Intelligence*
AD	Armazém de dados
OLTP	*Online Transaction Processing*
OLAP	*Online Analytical Processing*
SQL	*Structured Query Language*
CRUD	*Create Read Update and Delete*
ER	*Entity-Relationship modeling*
SCD	*Slowly Changing Dimension*
ETL	*Extract, Transform, Load*
SGBD	Sistema de Gestão de Base de Dados
KPI	*Key Performance Indicator*
ECTS	*European Credit Transfer and Accumulation System*
CIF	*Corporate Information Factory*
SSIS	*SQL Server Integration Services*
SSAS	*SQL Server Analysis Services*
MSSDT	*Microsoft SQL Server Data Tools*
DQP	*Data Quality Problem*
TFS	*Team Foundation Server*
HTML	*Hyper Text Markup Language*
LEI	Licenciatura em Engenharia Informática
ENAEE	*European Network for Accreditaion of Engineering Education*
MEI	Mestrado em Engenharia Informática

ROLAP	*Relational Online Analytical Processing*
MOLAP	*Multidimensional Online Analytical Processing*
HOLAP	*Hybrid Online Analytical Processing*
XML	*Extensible Markup Language*
ERP	*Enterprise Resource Planning*
CRM	*Customer Relationship Management*
LDAP	*Lightweight Directory Access Protocol*
ISO	*International Organization for Standardization*
CSAP	*Comprehensive Student Assistance Process*
A3ES	Agência de Avaliação e Acreditação do Ensino Superior

1 Introdução

Neste capítulo será descrito o objeto de estudo realizado no âmbito da unidade curricular Tese de Mestrado de Engenharia Informática (MEI), lecionada no Instituto Superior de Engenharia do Porto. Numa primeira abordagem será apresentado o problema encontrado atualmente, o que se pretende com este trabalho e que tipo de solução se adapta às necessidades encontradas, definindo os objetivos específicos do âmbito. Em forma de conclusão do capítulo, será descrita a estrutura seguida ao longo de todo o documento.

1.1 Enquadramento

Ano após ano, os dados curriculares de alunos que frequentam a Licenciatura em Engenharia Informática (LEI), do Instituto Superior de Engenharia do Porto, crescem e com isto cresce proporcionalmente o histórico académico de anos anteriores. Quando se fala em dados curriculares, refere-se às notas dos alunos em cada ano letivo, a uma determinada disciplina, em cada tipo de avaliação existente para a disciplina, entre outros dados. Todo este conjunto de informação é gerido pelo Portal da Instituição, tendo como base um sistema de armazenamento de dados (cujo âmbito é proprietário e desconhecido).

A direção de curso possui a necessidade de aceder e de conhecer estes dados académicos essencialmente para alunos que frequentam ou frequentaram a LEI, com o objetivo de construir relatórios, tornando a extração do conhecimento mais flexível e rápida. Atualmente, o diretor possui a responsabilidade de exportar a informação para folhas de cálculo, dado que a manipulação dos dados não pode ser feita através de ligação direta ao sistema, tornando-se um processo exclusivamente manual. Neste âmbito, torna-se possível reconhecer determinados comportamentos que fogem do que é espectável e definir um plano de ação em conformidade com as análises efetuadas aos dados escolares. Contudo, com este processo, as análises efetuadas dependem sempre da informação exportada, do volume de registos e do tipo de ferramentas que o formato aplicacional disponibiliza, sendo elevada a possibilidade de existência de erros entre registos e cálculos.

1.2 Motivação

Desta forma, surgiu a necessidade de criar uma solução que resolva as dificuldades encontradas, proporcionando análises de dados coerentes, fiáveis, orientadas ao negócio e sobretudo fáceis de criar e de utilizar, tendo sempre como base as fontes de dados de avaliações disponibilizadas.

Também se pretende que a performance de manutenção dos dados seja potenciada, através do carregamento de dados para um sistema centralizado e a utilização de ferramentas para extração e limpeza de dados automática.

1.3 Objetivos

Tendo em conta as necessidades enunciadas, a motivação e o contexto de avaliações em que se insere, foram definidos os seguintes objetivos:

- Criar um armazém de dados desenhado de acordo com as especificidades da instituição, nomeadamente através da identificação das principais métricas;
- Incluir neste armazém lógica associada às avaliações dos alunos;
- Montar o sistema de transformação, limpeza e carregamento dos dados, tendo em conta o tipo de formato das fontes de dados disponibilizadas e as características necessárias para a respetiva manutenção de histórico;
- Testar o funcionamento do processo de extração, transformação e carregamento dos dados;
- Disponibilizar as principais análises de avaliações de alunos utilizadas pelo Diretor de curso da LEI.

1.4 Estrutura do Documento

Primeiramente este documento começa pela introdução, onde é feita uma breve descrição do contexto e do problema que deu origem a esta dissertação. De seguida, no Capítulo 2 (Estado de Arte) são apresentados alguns conceitos da área, diferentes ferramentas de processamento e análise de dados, soluções existentes no mercado e indicadores utilizados no Ensino Superior Português.

No Capítulo 3 (Análise e Desenho), por forma a descrever a solução numa ótica mais tecnológica, é apresentado o capítulo de Análise e Desenho, onde são descritos todos os detalhes da

arquitetura da solução a implementar com base no problema apresentado. No Capítulo 4 (Implementação), é descrito ao pormenor o processo de implementação, descrevendo as instalações efetuadas para montar o ambiente com as ferramentas necessárias ao desenvolvimento e os passos de implementação do processo de carregamento de dados para cada uma das áreas. Em concordância segue-se o Capítulo 5 (Análise de resultados), onde é feita uma breve análise sobre a informação armazenada na solução através da criação de um conjunto de análises, com o objetivo de comparação com análises originais, podendo assim concluir como o sistema se comporta.

Por último, no Capítulo 6 (Conclusão) é apresentado o resumo da solução implementada, os respetivos benefícios e sugestões de melhorias futuras.

É de salientar que este documento se encontra estruturado de acordo com as regras de escrita técnico-científicas (Pereira 2016).

2 Estado de Arte

Este capítulo refere-se ao estado da arte, a base que sustenta todo o trabalho desenvolvido. No decorrer deste capítulo será apresentado todo o estudo efetuado, desde o conhecimento de arquiteturas e técnicas de modelação, à escolha da tecnologia certa para a concretização do armazém de dados (AD), incluindo ferramentas para recolha e análise de dados. Os critérios utilizados para a escolha têm como base o suporte oferecido, *standards* seguidos, bem como funcionalidades oferecidas, documentação e licenças. Performance e estabilidade da ferramenta não serão contabilizados pois isso requeria a implementação da solução pretendida nas várias ferramentas que serão expostas e, para além de não ser o âmbito desta dissertação, estas necessitam de ter licenças para serem utilizadas de forma livre e com acesso a todas as funcionalidades necessárias.

De notar que este capítulo foi desenvolvido com a colaboração do colega Tiago José Matos Pereira, aluno do MEI e igualmente do ramo de Sistemas Computacionais. Desenvolveu em paralelo um *data mart* relativo aos dados de inscrições de alunos.

2.1 Business Intelligence

> *"[…] o negócio é um conjunto de atividades realizadas por qualquer fim, seja ele ciência, tecnologia, comércio, indústria, lei, governo, defesa, etc. […] A noção de inteligência é definida também aqui, num sentido mais geral, como "a capacidade de apreender a inter-relação dos factos apresentados de modo a orientar a ação para o objetivo desejado.""* (Traduzido de (Luhn 1958)).

É possível definir BI (*Business Intelligence*) como um conjunto de técnicas e ferramentas usadas sobre grandes volumes de dados com o objetivo de obter conhecimento sobre o negócio em

questão, através de análises históricas e correntes sobre os dados. Atualmente existem diversas metodologias que permitem recolher dados de sistemas internos/externos a uma organização para posteriormente armazená-los, prepará-los para análise e assim criar relatórios capazes de evidenciar ao utilizador os principais indicadores de que este pretende, sem que conheça toda a arquitetura técnica que tem por base todo este mecanismo (Rouse 2015). Este tipo de análises são construídas com base nos dados previamente carregados no AD, onde o utilizador é livre de utilizar o tipo de métricas existentes, bem como o tipo de informação a ter em conta.

Uma das grandes vantagens deste processo é a rapidez com que os resultados são calculados e facilmente partilhados. Estes sistemas estão preparados para receber e processar grandes quantidades de dados, tornando por si só a tomada de decisões mais facilitada e rápida na medida em que o utilizador escolhe que tipo de análises pretende fazer aos dados. Depois de obtidos os dados necessários e de tomadas as decisões, torna-se possível ter uma visão concreta e fiável sobre os seus fundamentos, na medida em que todos os dados que entram no sistema sofreram um tratamento prévio. Pode ser considerada uma ferramenta muito útil para os responsáveis de empresas, permitindo uma rápida evolução não só a nível de tomada de decisões futuras como foi descrito, mas também a nível de análise do comportamento dos seus potenciais concorrentes e clientes, identificando possíveis melhorias nos produtos e segmentos de mercado que ainda não foram explorados pela mesma. Na Figura 1 é apresentado um resumo dos benefícios descritos, que facilitam as operações do quotidiano.

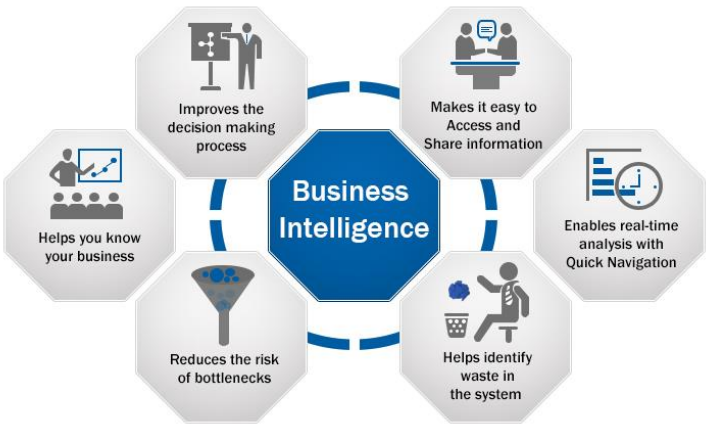


Figura 1 – Aplicação e vantagem na utilização de ferramentas inteligentes (Dean 2015).

2.2 Armazém de dados

De uma forma geral, um AD pode-se definir como um repositório de dados que passaram por um processo prévio de tratamento. Segundo *Bill Inmon*, um AD é o coração de todo o sistema onde está inserido, que disponibiliza informação para análise com o objetivo de servir de ferramenta de suporte na tomada de decisões. Considera que é um repositório de dados centralizados para toda a organização, caracterizando-o como orientado ao assunto do contexto onde está inserido, integrado devido ao facto de disponibilizar várias estruturas de

informação dentro do mesmo contexto, não volátil e variante no tempo na medida que não defende que existam atualizações diretas sobre os registos para que se consiga ver as diferenças ao longo do tempo. Para esse efeito é sempre adicionado um registo novo relativo ao momento temporal (Inmon 2002).

Já *Ralph Kimball* também o considera como ferramenta de suporte a tomadas a decisões. Deve ser um repositório facilmente acessível com dados tratados, consistentes e facilmente adaptáveis às mudanças permitindo a atualização direta/indireta de registos. Um AD pode ser composto por vários assuntos, sendo que cada área tem o seu processo de negócio distinto e que faz parte do contexto onde está inserido (Kimball & Ross 2013).

2.3 Armazenamento de dados – Armazém de dados versus Sistema Operacional

No contexto mencionado na secção anterior, depois de extraídos os dados, é necessário armazená-los em sistemas de repositórios de dados especializados, permitindo que possam ser preservados e consumidos *a posteriori*. Dependendo do tipo de características existentes e do contexto em que se insere, a estrutura do armazenamento pode ser distinta. Generalizando, existem dois tipos de sistemas: sistemas orientados à transação (sistemas OLTP - *Online Transaction Processing*) e sistemas orientados a análise/assunto (sistemas OLAP - *Online Analytical Processing*) (datawarehouse4u 2009).

Os sistemas operacionais (OLTP) têm como base uma arquitetura orientada à transação, isto é, tem como objetivo registar todas as transações efetuadas num determinado momento e em cada domínio. As bases de dados possuem uma estrutura relacional normalizada, onde os dados são armazenados em tabelas de acordo com o contexto e relacionadas entre os restantes artefactos existentes, criando dependências em rede. Dado o número máximo de dependências que pode existir em determinado contexto e a quantidade de dados adjacente, este tipo de sistemas pode perder a performance na consulta desses mesmos dados (Editorial Team+ 2007).

Por outro lado, o segundo sistema mais utilizado quando existe a necessidade de grande armazenamento de dados e respetiva análise é o AD. É um sistema computacional capaz de armazenar grandes quantidades de dados e manter a performance nas consultas devido à sua estrutura desnormalizada e com poucas dependências entre domínios de dados. Armazena todo o conjunto de dados em modelos multidimensionais/dimensionais, constituídos por dimensões e tabelas de factos. Neste tipo de modelo, as dimensões armazenam todos os dados que pertencem ao seu domínio. Como exemplo deste tipo de armazenamento temos o caso do aluno, onde seria especificada uma estrutura capaz de armazenar os dados básicos de cada aluno existente. No que toca às tabelas de factos, estas são responsáveis por relacionar os

registos existentes em cada domínio, armazenando acontecimentos e KPIs [1] previamente determinados na fase de análise do sistema (Ballard et al. 2006).

Assim, por sistema OLAP pode definir-se um conjunto estruturas de informação criadas a partir do modelo dimensional do AD, onde podem conter a informação armazenada de forma agregada, contabilizando um conjunto de factos e medidas que se pretende de acordo com o contexto. Em alguns casos, as consultas de dados passam a ser feitas diretamente sobre este tipo de estruturas, libertando alguma carga sobre o AD. No capítulo *Online Analytical Processing* é descrito com maior nível de detalhe como funcionam este tipo de metodologias.

Cada um dos tipos de armazenamento enunciados possuem as suas especificidades. Devido ao processo prévio de limpeza e tratamento de dados (ETL - *Extract, Transform, Load*), o AD permite despistar em grande alcance as inconsistências, bem como definir qual a melhor resolução do problema perante o caso em questão. No sistema operacional, o utilizador é que necessita de tomar a iniciativa de limpeza das inconsistências de forma manual. Não possui um processo estruturado que faça parte de um possível carregamento propriamente dito, ficando assim ao critério do criador do sistema aquando ou depois da introdução dos dados no sistema operacional (Tech-FAQ 2013).

O acesso à informação nas análises pode atingir tempos de resposta muito expectantes, ao contrário do que acontece com os sistemas operacionais. No AD e como já foi referido, as consultas efetuadas pelo consumidor podem não ser efetuadas diretamente sobre o sistema de armazenamento. Depois do sistema se encontrar limpo e carregado, é criada uma camada de abstração para que todas as consultas incidam sobre a camada. No caso dos sistemas operacionais, todas as interrogações e consultas que se efetuam ao sistema, são inteiramente efetuadas diretamente.

De notar que, apesar da execução de análises sobre o AD obter bons resultados, a desvantagem aparece no carregamento/tratamento de dados para este sistema e respetiva manutenção, sendo um processo lento e trabalhoso (Tech-FAQ 2013). O AD permite que o carregamento de dados seja efetuado a partir de fontes de dados de tipos diferentes, nomeadamente ficheiros Excel, ficheiros de texto, bases de dados, etc. No entanto e tendo em conta as desvantagens enunciadas relativamente aos sistemas operacionais, é de notar que estes são normalmente usados como fontes de dados para carregar o AD (Ballard et al. 2006).

Para implementar um sistema de armazenamento, é necessário ter em conta as necessidades do meio e o benefício/custo que cada tipo de solução terá e escolher qual a que melhor se ajusta ao caso.

[1] Conjunto de indicadores chave que permitem compreender se os objetivos definidos estão a ser cumpridos, medindo o desempenho, o sucesso ou a falha (IBM 2015).

2.4 Online Analytical Processing

OLAP (*Online Analytical Processing*) é o mecanismo de análise de sistemas multidimensionais que disponibiliza a capacidade de realizar operações complexas e sofisticadas sobre este tipo de modelos. Permite aos utilizadores finais realizar pesquisas em múltiplas dimensões em simultâneo, disponibilizando a informação de que necessitam para tomarem as suas decisões. A vantagem de utilizar OLAP está na velocidade de acesso à informação armazenada no modelo multidimensional, criando agregações e cálculos muito rapidamente em múltiplos conjuntos de dados. A implementação deste tipo de estrutura depende não só do tipo de *software* que se está a utilizar mas também do tipo das fontes de dados e dos objetivos do negócio em que se insere (OLAP.com 2016).

A estrutura deste tipo de metodologia tem como base um cubo de dados dividido em pequenos segmentos (que representam determinado valor), em que cada vértice é referente a um domínio de informação específico (dimensão). Na Figura 2 é apresentado o exemplo de um cubo, que relaciona a informação da região, do produto e do dia em que foi adquirido (Microsoft 2016b).

Figura 2 – Cubo OLAP (Microsoft 2016b).

Para obter a informação pretendida, existem um conjunto de operações que podem ser feitas sobre esta estrutura de dados que consistem em combinar os valores das dimensões, atuando como filtro para obter a informação necessária. As operações são as seguintes (em contexto com o cubo apresentado na Figura 2)(tutorialspoint 2016; Microsoft 2016b):

- ***Drill-down:*** este tipo de operação baseia-se numa procura de dados progressivamente afunilada, elevando-se a granularidade da pesquisa para obter informação cada vez mais detalhada a cada nova procura efetuada. Um exemplo seria progredir de uma pesquisa de vendas de dispositivos móveis por períodos de tempo diferentes: no ano corrente, nos primeiros cinco meses do ano, num mês em específico, numa semana ou até no dia corrente;

- ***Roll-up*****:** descreve-se como sendo a operação contrária à anteriormente enunciada. Parte-se de um nível de granularidade mais específico, diminuindo-a progressivamente. Neste contexto, o filtro começaria pelo dia, seguido do mês e terminaria no ano;

- ***Drill-across:*** tipo de operação que incide em dados que tem como origem mais do que uma tabela de factos e que se relacionam a partir das ligações às dimensões em comum.

Um exemplo seria a existência da tabela de factos para o processo de vendas de produtos e uma outra tabela de factos para o processo de queixas associadas a determinado produto. Teriam as três dimensões presentes como comuns, dado que a queixa é registada em determinado dia, é associado a determinado produto (que foi adquirido por outrem) e numa determinada região. Resumindo, poderia ser feita uma pesquisa para obter o conjunto de queixas de um determinado produto que foi adquirido em 2016;

- ***Slice:*** este tipo de operação cria um novo cubo de informação segmentado, baseando-se em apenas um filtro de uma das dimensões. Ao pesquisar por todas as vendas de produtos da categoria dispositivos móveis, é obtido um conjunto de informação específico com apenas um filtro;

- ***Dice:*** nesta operação também é criado um novo cubo à semelhança da operação descrita anteriormente, à exceção de que são utilizados dois ou mais filtros. Um exemplo deste tipo de operação seria pesquisar por todas as vendas de produtos do tipo dispositivos móveis em 2016;

- ***Rotation:*** baseia-se na rotação dos eixos dos cubos, com o objetivo de permitir elaborar pesquisas em diferentes perspetivas.

Existem três tipos de estruturas distintas para o armazenamento dos dados, denominadas por MOLAP, ROLAP e HOLAP. No MOLAP, (*Multidimensional Online Analytical Processing*) os dados estão armazenados num cubo multidimensional, proporcionando a rapidez na obtenção dos dados e nas operações de *slicing* e *dicing* em cubos. Por sua vez, o ROLAP (*Relational Online Analytical Processing*) consiste em efetuar análises em dados armazenados num modelo relacional e efetuar as operações como no MOLAP, com a diferença de que todas as interrogações são feitas diretamente ao AD. Por fim, o HOLAP (*Hybrid Online Analytical Processing*) consiste na junção do MOLAP e ROLAP, sendo que o HOLAP aproveita a tecnologia do cubo para um processamento mais rápido (1keydata 2015).

Na Tabela 1 encontram-se enunciadas as principais diferenças entre um sistema OLAP e um sistema OLTP.

Tabela 1- OLAP versus OLTP (Atanazio 2013).

	OLAP	**OLTP**
Fontes de dados	Os dados provém dos sistemas do tipo dimensional e/ou relacional.	Sistemas operacionais
Propósito dos dados	Ajudar na tomada de decisão e planeamento	Controlar e executar as operações de negócio
Queries	Complexas	Simples
Velocidade de processamento	Depende do volume de dados a analisar	Rápido Processamento
Requisitos de espaço	Grande capacidade	Relativamente pequeno
Modelo da Base de Dados	Multidimensional	Relacional
***Backup* e Recuperação**	Backup não regular	Backup regular
Idade dos dados	Histórico	Corrente
Operações efetuadas	Ler	Adicionar, atualizar, ler e eliminar
O que os dados revelam	Vistas multidimensionais de vários tipos de atividades de negócio	Imagem do processo de negócio corrente

2.5 Data Mart

Um *Data Mart* é um sistema de divisão lógica mais pequeno que fornece suporte à tomada de decisões para uma determinada área de negócio em específico (por exemplo: Vendas, *Marketing*, etc.). O AD pode ser dividido/composto por várias áreas deste formato, tornando-se mais fácil de gerir e manter na medida em que as operações necessárias a serem efetuadas apenas incidem num determinado domínio, mantendo os restantes operacionais e com impacto reduzido (Ballard et al. 2006). A Figura 3 representa um exemplo de quatro *data marts*, sendo que cada um possuí um âmbito distinto: armazenamento de informação relativa a vendas de produtos, dados dos clientes, dados das compras e armazenamento de informação relativamente ao inventário de cada uma das lojas.

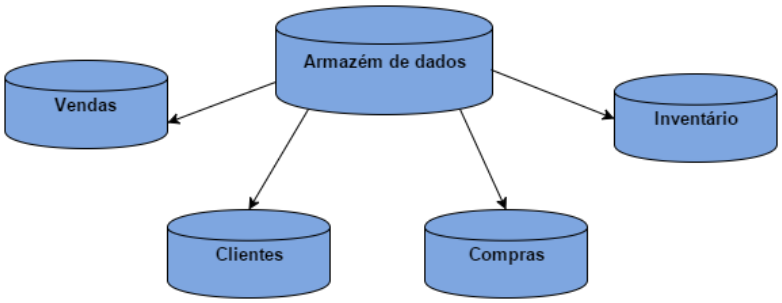


Figura 3 – Exemplo de um AD composto por *data marts*.

Estes podem ser classificados como dependentes ou independentes de acordo com o seu nível de dependência. São dependentes quando estes são parte integrante do AD, ou seja, significa

que a estrutura do armazém é definida por todos os *data marts* nele contidos. O contrário aplica-se quando o AD possui a sua própria lógica e estrutura, sendo que apenas possui ligações com *data marts* externos, sendo assim cada um deles independente do contexto do AD. (Oracle 2007). Na secção seguinte é possível ver com mais detalhe cada um dos tipos enunciados.

2.6 Arquiteturas de armazéns de dados

Cada AD desenvolvido possui uma estrutura distinta, com características específicas e relacionadas com o meio em que está inserido. Nesta secção serão abordadas as principais teorias existentes atualmente, defendidas por *Bill Inmon* e *Ralph Kimball*.

2.6.1 Arquitetura *Ralph Kimball*

> *"Pense como um restaurante. Imagine que os clientes do restaurante são os utilizadores finais e a comida são os dados. Quando os alimentos são oferecidos a todos os clientes na sala de jantar, estes são servidos exatamente no local e na forma como esperam receber: limpos, organizados e apresentados de uma forma em que cada peça pode ser facilmente distinguida e consumida. […] Na cozinha, a comida é selecionada, limpa, cortada, cozinhada e preparada para apresentação."* (Traduzido de (Kimball & Ross 2013)).

Segundo *Ralph Kimball*, um AD encontra-se dividido em duas áreas lógicas, podendo em alguns casos encontrarem-se separadas fisicamente: camada de implementação (*The Back Room)* e a camada de apresentação (*The Front Room)*. De uma forma breve, na primeira área encontra-se toda a lógica criada incluindo ligações às fontes de dados (*Source Systems*), processos para extração e tratamento desses mesmos dados e processos para armazenamento num sistema de base de dados de *staging* (*The Staging Area*). Esta secção também inclui AD, que vai ser carregado a partir dos dados resultantes do processo de tratamento de dados.

Até aqui, o acesso é interdito por parte da camada de apresentação, sendo que os dados passam a estar disponíveis para os utilizadores finais através da camada de apresentação que engloba todas as aplicações de suporte ao uso e análise dos dados (*The Presentation Area*) (Kimball & Ross 2013). Retirada da mesma fonte de informação, na Figura 4 é apresentada uma ilustração da composição das divisões lógicas descritas.

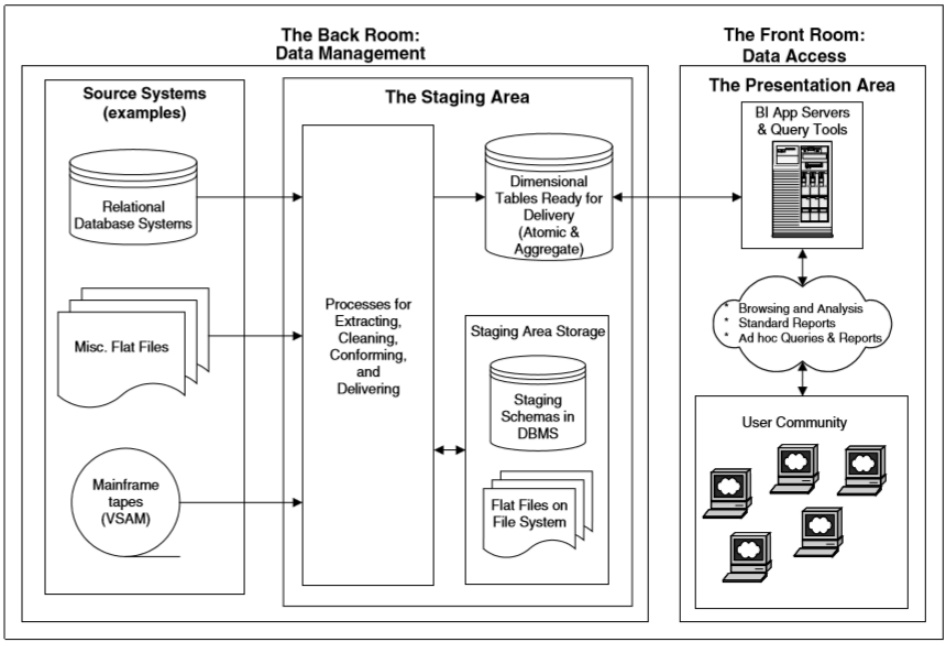


Figura 4 – Conceito camada de apresentação *versus* camada implementação (Kimball & Caserta, 2004).

Tomando como foco a primeira camada, todo o processo começa a partir das fontes de dados. Estas fontes são as que possuem os dados necessários que *a posteriori* vão alimentar o AD podendo assumir formatos variados desde uma base de dados operacional, atualmente utilizada para a gestão do negócio, até a ficheiros de texto. Primeiramente é feita uma extração dos dados em bruto, de uma ou mais fontes, podendo ser armazenada previamente facilitando o recomeço do processo de carregamento sem ter a necessidade que sobrecarregar novamente o sistema na extração seguinte. Depois de extraídos os dados, estes passam para uma área denominada por *Staging Area* (Kimball & Caserta 2004; IBM 1999).

> *"A cozinha é uma área de trabalho, fora do alcance dos clientes do restaurante."* (Traduzido de (Kimball & Caserta, 2004)).

Continuando com a analogia, este define-a como uma área de limpeza e tratamento dos dados, sendo composta por um processo de extração, tratamento, carregamento (processo ETL) e por um sistema de armazenamento interno invisível para o utilizador, com o fim de armazenar os dados tratados antes de passar para o AD final. Após estes dados serem carregados em *staging*, recebem o último tratamento antes do armazenamento no modelo multidimensional, denominado por *Enterprise Bus Architecture*. Estes dados encontram-se na forma mais atómica possível para permitir dar resposta quer a query's imprevisíveis, bem como a agregações (*roll-up*) ou conseguir ir a um maior nível de detalhe (*drill-down*) (Oracle 2007). Depois de estes dados serem inseridos no AD, podem ser feitas manutenções a nível de histórico e atualizações aos registos.

Esta arquitetura defende uma metodologia do tipo *bottom-up*, considerando que a criação de valor para o AD pode ser incrementada ao longo do tempo através de novos *data marts*, acrescentando assim novos domínios de informação. Não requer que todos estes domínios sejam criados logo na fase inicial. A Figura 5 apresenta de forma resumida, a estrutura global de um sistema construído com base nesta metodologia.

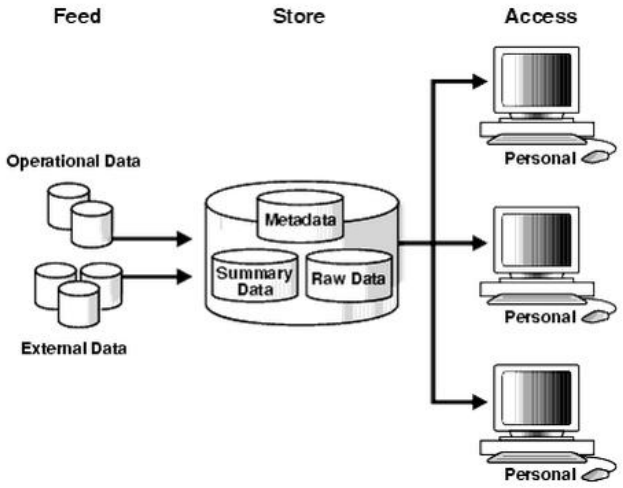


Figura 5 - Arquitetura de *Ralph Kimball* (Oracle 2002).

2.6.2 Arquitetura *Bill Inmon*

> *"O armazém de dados é o coração do ambiente arquitetado, e é o fundamento de todo o processamento […] Existe uma única fonte integrada de dados […] está orientado para as principais áreas de negócio que foram definidas no modelo de dados corporativo de alto nível. Cada área de negócio está fisicamente implementada como uma série de tabelas relacionadas no armazém de dados."* (Traduzido de (Inmon 2002)).

Inmon defende que todos os dados operacionais que existem numa organização são parte integrante do AD. Este tipo de arquitetura assenta numa metodologia do tipo *top-down, onde* primeiramente se procura identificar todos os domínios de dados existentes com o objetivo de criar relações lógicas entre si, resultando assim num modelo de dados essencialmente relacional. Desta forma, a estrutura total do AD é definida e só depois possivelmente se dividirá em diferentes *data marts*, incluindo toda a estrutura numa área denominada por CIF (*Corporate Information Factory*).

> *"Um armazém de dados é um conjunto de dados de suporte a decisões de administração orientado ao assunto, integrado, não volátil e variante no tempo"* (Traduzido de (Inmon 2002)).

Este tipo de arquitetura possui algumas especificidades, distintas da arquitetura apresentada anteriormente. O AD é orientado ao assunto, ou seja, a informação encontra-se organizada de acordo com as relações que possui, resultando numa base de dados operacional normalizada

na terceira forma normal. Estas relações existem dado que esta arquitetura defende que todos os dados que possuam o mesmo domínio de informação devem ser relacionados. Relativamente à atualização dos dados, não existe qualquer possibilidade de o fazer. Cada um dos registos é preservado e guardado tal e qual como foi inserido para efeitos de análise futuras, registando apenas o momento de inserção para existir a perceção da variável tempo perante os restantes dados. Por último, este tipo de AD é também integrado dado que pode conter informação de mais do que um sistema relacional existente, tornando a informação nele contida consistente por seguir uma metodologia igualmente relacional (Oracle 2002).

Como apresenta a Figura 6, a nível de extração, tratamento e carregamento de dados, é igualmente utilizada uma área de *staging* como a primeira camada a receber informação das diferentes fontes de dados. Toda a informação passa desta área diretamente para o AD, sendo a partir deste sistema que cada *data mart* de dados é logicamente criado e carregado.

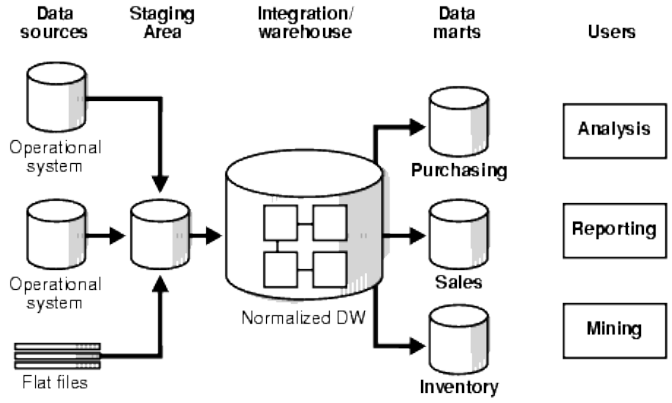


Figura 6 – Arquitetura de *Bill's Inmon* (Oracle 2002).

2.6.3 Comparação entre arquiteturas

Na Tabela 2, é feita uma comparação entre as duas arquiteturas enunciadas, apresentando os pontos positivos e negativos em cada uma.

Tabela 2 – Comparação entre arquiteturas *Kimball* e *Inmon* (Inmon 2002; Kimball & Caserta 2004; Kimball & Ross 2013; Sansu George 2012).

	Arquitetura *Ralph Kimball*	**Arquitetura *Bill Inmon***
Dados de negócio e o AD	O AD é composto por vários *data marts* em que cada um é responsável por um segmento no global do negócio. Todos os *data marts* resultam assim no AD.	Todos os dados são parte integrante do AD. É definida uma estrutura global e só depois, se existir essa necessidade, podem ser criados segmentos à parte.
Staging area	Defende o conceito.	Defende o conceito.
ETL	Defende o conceito.	Defende o conceito.
Data Marts	Numa primeira fase, são definidos cada um dos segmentos, dando origem a *data marts* distintos. Só depois é que o AD é definido (abordagem *bottom-up*).	Logo à partida é definida a estrutura do AD. Apenas se necessário, esta estrutura pode ser segmentada em *data marts* distintos (abordagem *top-down*).
Variante no tempo	Defende o conceito.	Defende o conceito.
Modelo de AD	Segue o modelo multidimensional.	Essencialmente relacional (na terceira forma normal).
Orientação aos processos	Não, é orientado ao assunto.	Sim.
Complexidade de desenvolvimento	Simples, na medida em que cada estrutura de dados existente é pensada e segmentada logo na fase da construção do AD. O tipo de relações que este tipo de estruturas pode ter é favorável, dado que as dependências são poucas, baseando-se no geral em relações entre dimensões e tabelas de factos.	Complexa, dado que primeiramente se constrói uma estrutura global para todo o tipo de informação da organização. As relações entre dados podem ser de perceção complexa e pode deteorar a performance a nível de pesquisas de dados se a dependência entre dados for muito grande.
Registo de alterações nos dados	Suporta. É a arquitetura defensora.	Não defende.
Tempo de desenvolvimento	Menor tempo de desenvolvimento.	Maior tempo de desenvolvimento.
Custo	Menor custo de desenvolvimento inicial. Cada segmento que seja construído numa fase posterior terá exatamente o mesmo custo.	Maior esforço inicial. Os desenvolvimentos seguintes terão menor custo de desenvolvimento.
Conhecimentos requeridos	Não são requeridos conhecimentos especialistas, apenas generalistas.	Equipa especializada, dada a complexidade do modelo.

Concluindo, cada uma das arquiteturas possui as suas especificidades. Aquando da escolha da melhor arquitetura deve ser tido em linha de conta o contexto e as necessidades por satisfazer,

optando por uma solução progressiva de *Kimball* ou uma solução completa na fase de criação de *Inmon*.

2.7 Modelação Dimensional

Atualmente os modelos multidimensionais constituem uma base sólida de armazenamento e de gestão de dados nas soluções de BI (Elias 2015). Estes modelos permitem a definição do relacionamento dos dados, concebendo um suporte a consultas em todas as vertentes de negócio, bem como a extração de detalhes sobre esses mesmos dados. Desta forma torna-se mais fácil compreender toda a diversidade de informação de uma organização, de forma mais intuitiva e eficaz.

Este tipo de modelo segue uma estrutura de dados e uma forma de desenvolvimento do sistema de armazenamento muito específica, distinguindo-se claramente do modelo relacional baseado no cumprimento das formas normais com o objetivo da obtenção de um modelo totalmente normalizado. A modelação dimensional defende assim existência de uma ou mais tabelas que registam todas as ocorrências na forma mais resumida possível (tabela de factos), criando tabelas distintas associadas que especificam cada um dos domínios de informação (tabela de dimensão). É necessário assim racionalizar logicamente, identificando o que é necessário registar em cada evento e a informação detalhada que pode ser aglomerada em tabelas diferentes, permitindo assim que todos os registos que possam vir a ser armazenados se possam relacionar com outros eventos.

> *"A tabela de factos contém medidas numéricas resultantes das medidas operacionais utilizadas no mundo real. Com menor granularidade, uma linha da tabela de factos corresponde a um evento e vice-versa. Assim, a estrutura fundamental de uma tabela de factos é inteiramente baseada em acontecimentos e não pelos eventuais relatórios de informação."* (Traduzido de (Kimball & Ross 2013)).

As tabelas de factos, como o próprio nome indica, têm como objetivo armazenar registos associados a operações ou eventos de um determinado contexto previamente identificado. Assim, para cada evento são definidos os campos necessários que o descrevem e a granularidade pretendida, identificando o que se pretende medir numericamente com a informação e qual o nível de detalhe pretendido.

Na Figura 7 é apresentado um exemplo de como construir a estrutura com os campos necessários de forma a identificar cada evento no contexto da venda de produtos. Apesar do exemplo apresentado ser mais abrangente, a termos de um exemplo simples pretende-se medir as vendas efetuadas pelos clientes e analisar detalhadamente a quantidade de produtos comprados de um determinado tipo, contabilizando o preço total da compra, o número de produtos adquiridos e o peso.

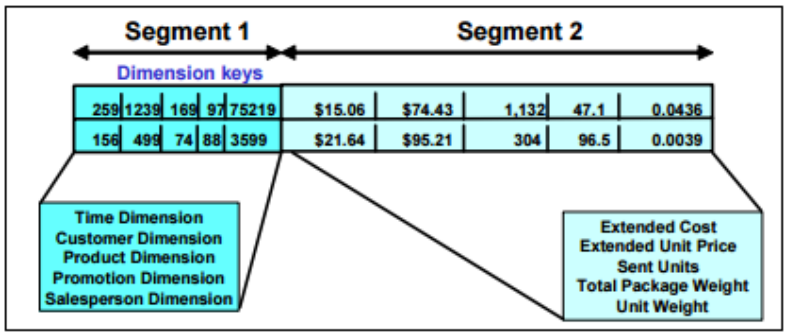


Figura 7 – Exemplo do pensamento lógico na definição da tabela de factos (Ballard et al. 2006).

É possível então definir que a tabela de factos necessita de armazenar informação sobre o cliente, os produtos adquiridos, a quantidade respetiva, o preço total da compra e o respetivo peso da compra. Depois de definida a base que sustenta a tabela, é então necessário identificar os elementos principais que estão relacionados com a venda: clientes e produtos. Toda a restante informação é contabilizada como medida para a tabela de factos, dado que cada compra possui o seu preço específico, a sua quantidade específica e peso dos produtos igualmente específico ("*Segment 2*").

No entanto, para cada elemento principal acima identificado, é analisada a necessidade de dividir a informação em vários domínios, permitindo assim relacionar o mesmo cliente com mais do que um ato de compra sem que toda a informação detalhada sobre o mesmo se repita e que seja identificada de uma forma simples ("*Segment 1*"). Estas estruturas relacionáveis denominam-se por tabelas de dimensão.

> *"Dimensões fornecem o contexto "quem, o quê, onde, quando, porquê e como" relacionado com um evento de todo o conjunto dos processos de negócio. As tabelas de dimensão contêm os atributos descritivos utilizados por aplicações BI para filtrar e agrupar os factos. […] Sempre que possível, uma dimensão deve identificar um domínio específico quando relacionado com uma linha da tabela de factos."* (traduzido de (Kimball & Ross 2013)).

Uma tabela de dimensão consiste assim, em adicionar um contexto aos eventos existentes numa tabela de factos. Ainda sobre o exemplo apresentado, o domínio cliente teria toda a informação detalhada sobre o mesmo desde o nome, idade, morada, código postal, número de identificação fiscal, entre outros tipos de informação possíveis. Este tipo de tabela contém uma relação com a tabela de factos através de uma chave natural (denominada *surrogate key*), onde cada chave identifica de forma unívoca o registo, relacionando-o com cada facto da tabela de factos (Figura 8). Este tipo de tabelas, como são desnormalizadas, possuem um conjunto de campos maior e com menos quantidade de registos armazenados.

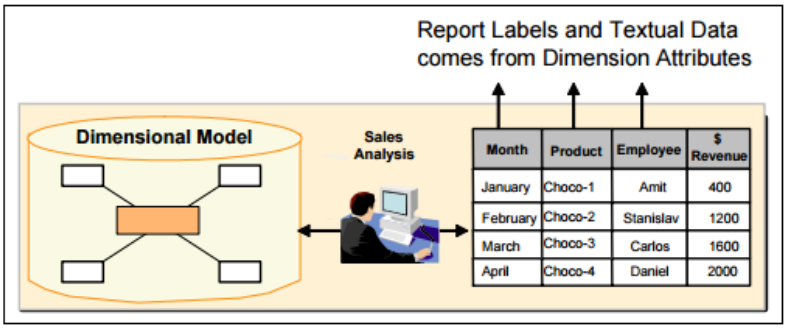


Figura 8 – Exemplo de análise de vendas, com base na informação existente nas dimensões relacionadas (Ballard et al. 2006).

À medida que cada um dos artefactos vai sendo criado, o modelo vai assumindo uma estrutura muito específica sobre o contexto em que se insere. Essa estrutura pode assentar num dos seguintes tipos de modelos possíveis (Figura 9):

- **Modelo em estrela (*star-schema*):** este modelo é composto por uma única tabela de factos e por um conjunto de dimensões todas relacionadas;

- **Modelo em floco de neve (*snowflake model*):** é composto por uma tabela de factos e por dimensões onde, para além de se relacionarem com a tabela de factos podem relacionar-se entre dimensões, acabando por seguir uma abordagem de normalização, mesmo que significativa. Um exemplo desta técnica é o exemplo da existência de uma dimensão cujo objetivo é armazenar a informação básica de cada estabelecimento onde são vendidos/comprados os produtos. É importante determinar qual a localização do estabelecimento e para isso é necessário armazenar um conjunto de campos: morada, código postal, entre outros. Esta informação pode ser convertida para uma nova dimensão denominada "Região", permitindo que uma mesma localização nela existente possa ser mapeada para mais do que um estabelecimento, sem que todo o detalhe relativo à localização se repita;

- **Multi-estrela (*multi-star model*):** modelo composto por mais do que uma tabela de factos, ligadas indiretamente através das dimensões que possuem ligação comum.

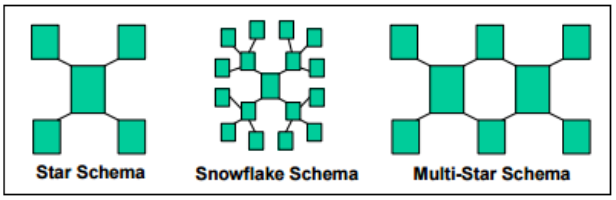


Figura 9 – Tipos de modelos dimensionais (Ballard et al. 2006).

Para cada tipo de tabela enunciada, existem estruturas ainda mais específicas que permitem resolver determinados problemas. Estas estruturas já foram pensadas por quem já passou por

essas dificuldades, assemelhando-se ao uso de padrões para resolução de alguns problemas. De seguida, vão ser apresentados alguns tipos de tabelas de factos e tipos de dimensões.

Tabelas de Factos

- **Tabelas de factos transacionais:** tabela que regista um conjunto de acontecimentos, em que cada linha representa um evento em específico;

- **Tabelas de factos agregadas (*Agregated fact table*):** este tipo de tabelas tem como base a informação armazenada na tabela de factos mas de uma forma agregada, definindo-se no carregamento os níveis de informação pretendidos. Tem grandes vantagens a nível de execução de consultas, dado que não é necessário reagregar a cada utilização e não impede que a informação seja novamente agregada;

- **Tabelas de factos acumulativas (*Cumulative Fact Table*):** estas tabelas têm como objetivo armazenar todos os registos desde que o processo de negócio arrancou até terminar. Cada registo representa um processo que é atualizado ao longo do tempo;

- **Tabelas de Factos periódicas (*Periodic Fact Table*):** este tipo de tabela de factos tem como granularidade principal o tempo. Contém factos proveniente de uma tabela de factos transacional, mas apenas sobre determinado período de tempo;

- ***Factless fact tables:*** Este tipo de tabelas de factos caracteriza-se por não serem constituídas por medidas. Apenas cruzam informação vinda de diferentes dimensões, através da chave de relacionamento.

Tabelas de Dimensão

- **Dimensão conforme (*Conformed dimensions*):** Tem como objetivo agrupar um conjunto de informação sobre um domínio específico, podendo este tipo de estruturas ter relacionamentos com mais do que uma tabela de factos ao mesmo tempo e em *data marts* diferentes. Tomando como exemplo uma dimensão responsável por armazenar informação de variável temporal, a dimensão Data, que armazena campos como a data completa, o dia, mês, ano, número do dia do ano, valor que identifica se é dia de trabalho, entre outros. Esta dimensão pode ser relacionada com outras através da sua chave que identifica de forma temporal qualquer entidade que esteja relacionada;

- **Role playing dimensions:** As dimensões tem esta denominação quando estão relacionadas com a mesma tabela de factos mais do que uma vez. Isto acontece, dado que para um domínio de informação podem existir diferentes significados a nível do contexto real, como por exemplo a utilização da dimensão Data. Um processo de venda pode conter vários tipos de datas diferentes: uma data da venda, data de vencimento, data de pagamento, entre outras. Assim, a tabela de factos possui mais do que um campo relacionado com a dimensão em questão, sendo cada um deles diferente e

constituído pela chave da dimensão em causa para a relação (proveniente da mesma dimensão);

- **Dimensão ponte (*Bridge table*):** Este tipo de dimensão atua como tabela de união entre duas outras dimensões. É utilizada quando estamos perante registos que possuem relações do tipo muitos-para-muitos (*many-to-many*), ou seja, quando os registos de uma tabela se relacionam com mais do que um registo da outra tabela e vice-versa. Um exemplo prático para este tipo de tabela é o contexto da banca. Um cliente tem uma conta associada e, por sua vez, cada conta pode estar associada a mais do que um cliente (diferentes titulares);

- **Junk Dimension:** Podem existir certos conjuntos de informação que, apesar de não formarem um único domínio em específico, repetem-se entre registos na tabela de factos e ao qual pode ser considerada a possibilidade de exportar esses padrões para uma dimensão à parte. O exemplo da Figura 10 apresenta uma tabela de factos cujo contexto relaciona-se com a entrega de encomendas.

  É possível observar que existem um conjunto de campos onde o valor se repete: confirmado ("*confirmed*), não confirmado (*not confirmed*), entregue (*delivered*), não entregue (*not delivered*), frágil (*fragile*). Todas estas combinações de dados podem ser armazenadas numa dimensão à parte, sendo apenas a sua chave a parte integrante da tabela de factos. Desta forma, a redundância a este nível é maioritariamente eliminada, sendo que a chave utilizada pode ser utilizada em qualquer facto e ao qual relaciona todo o detalhe necessário num local único, diminuindo o tamanho da tabela de factos;

Item	Client	QTY	Amount	Confirmed	Delivered	Fragile	Del. Method	Invoiced
100	A123	150	15000	Confirmed	Not Delivered	Yes	Standard	Not Invoiced
200	A123	341	76000	Not Confirmed	Not Delivered	No	Standard	Not Invoiced
100	B222	140	12500	Confirmed	Delivered	Yes	Express	Invoiced
100	C112	900	85000	Confirmed	Delivered	Yes	Express	Invoiced
200	C112	600	99060	Not Confirmed	Not Delivered	No	Standard	Not Invoiced

Figura 10 – Exemplo *junk dimension* (Jet Reports 2016).

- **Dimensão degenerada (*Degenerate dimension*):** Este conceito é dado a uma chave de dimensão utilizada em tabelas de factos mas ao qual não possui nenhuma dimensão física associada. Normalmente, esta técnica é usada quando se pretende armazenar, junto com o facto, a que conjunto de registos pertence. No exemplo da tabela de factos anterior que armazena linhas de encomenda, para cada linha existente seria adicionado um campo com o identificador da encomenda. Ao fazer a pesquisa pelo identificador, é retornado o conjunto de linhas de encomenda associado.

A técnica utilizada para o desenho de um sistema de armazenamento de dados prende-se em muito com o contexto em questão. Não é objetivo possuir estruturas completamente

relacionadas e com uma grande quantidade de dados/registos, nem o contrário. O equilíbrio determina a performance que se pode obter na camada de consulta e apresentação de dados.

2.8 Ferramentas ETL

Para implementar um AD, é fulcral que a fase de ETL seja corretamente implementada e flexível. Para isso, a escolha de uma ferramenta para o processo é fundamental sendo que, a versatilidade e configurabilidade que se pode adicionar num processo de ETL é sempre bem vista com o objetivo que este sistema evolua e se adapte a diferentes cenários. Nos tópicos seguintes são apresentadas algumas ferramentas de ETL que estão no top dez de utilização pelas empresas (ETL 2015).

2.8.1 Oracle Data Integrator

O ODI é uma ferramenta desenvolvida pela *Oracle* que permite a integração, transformação, replicação, gestão de meta dados, serviços e qualidade de dados num AD. Recentemente, foi lançada uma versão que permite a integração com a *cloud* e com possibilidade de efetuar análises em soluções *Big Data*[2], sendo por isso uma mais-valia (Oracle 2016; ETL 2015).

Permite interagir com vários sistemas heterogéneos através da utilização de outras ferramentas complementares, aumentado a performance das soluções de BI. Como consequência da interação, através do *Oracle GoldenGate* é possível obter dados em tempo real, através de processos de sincronização de dados nestes sistemas. Desta forma, o desempenho é maximizado aquando da migração de dados, sem tempo de inatividade, possibilitando a recuperação em caso de desastre e respetiva sincronização ativa e contínua nas bases de dados. Esta ferramenta também possui componentes *drag and drop*, que abstraem da maior parte de implementação apenas dando pequenas instruções e comandos SQL (*Structured Query Language)* para efetuar o processo de ETL (Narasimharajan 2011). Para além dos benefícios, a documentação sobre a ferramenta é escassa para ajudar a integrar e a criar processos de carregamento e extração de dados, existindo por outro lado formações da *Oracle* mas com um custo associado para as adquirir (InformationWeek 2015).

Na Figura 11 é apresentado um exemplo do ambiente de utilização da ferramenta, na criação de um processo de extração e carregamento de dados relativa a clientes.

[2] Termo utilizado para descrever a existência de uma grande quantidade de informação.

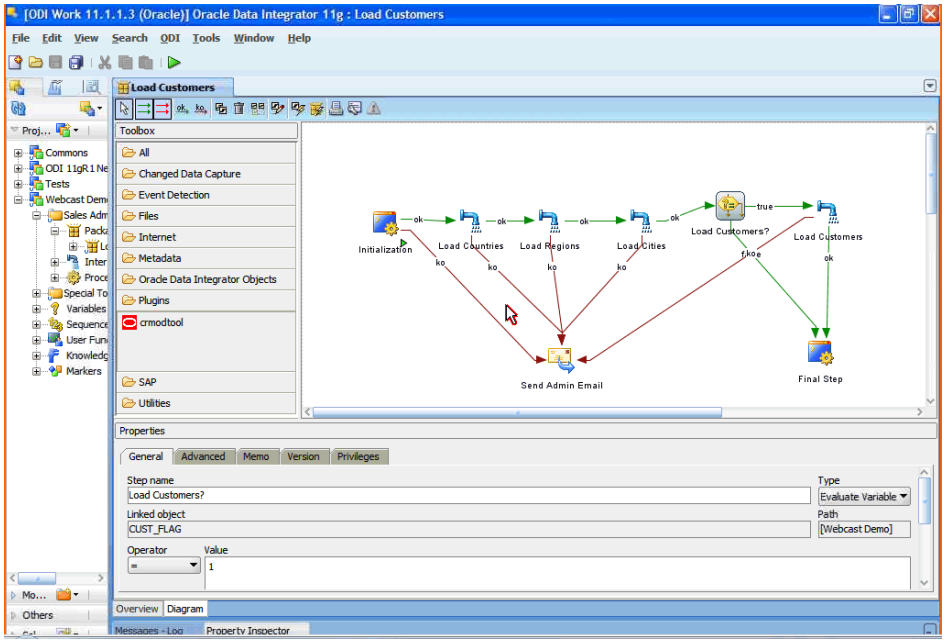


Figura 11 – Oracle ODI(ARSON Group SAC 2016).

2.8.2 Microsoft SQL Server Integration Services

O *Microsoft Integration Services*, mais conhecido por SSIS (*SQL Server Integration Services*), é uma plataforma de integração de dados utilizada a nível empresarial. Este tipo de ferramenta é utilizado para trabalhar soluções com alguma complexidade de lógica de negócio, permitindo o armazenamento, limpeza e gestão de dados. Auxilia na resolução de eventuais problemas, permitindo copiar ou descarregar ficheiros, enviar correio eletrónico com a informação de erros ou apenas a título informativo de que algum evento ocorreu. É de salientar que é possível utilizar pacotes que podem correr em simultâneo e suportam dados provenientes de várias fontes de dados (Microsoft 2015). A Figura 12 apresenta um exemplo breve de um fluxo de dados criado, onde duas fontes de dados distintas são integradas num único sistema de armazenamento, definindo qual o critério de união entre registos (componente *merge join*) e quais os registos do conjunto que são efetivamente válidos (componente *conditional split*).

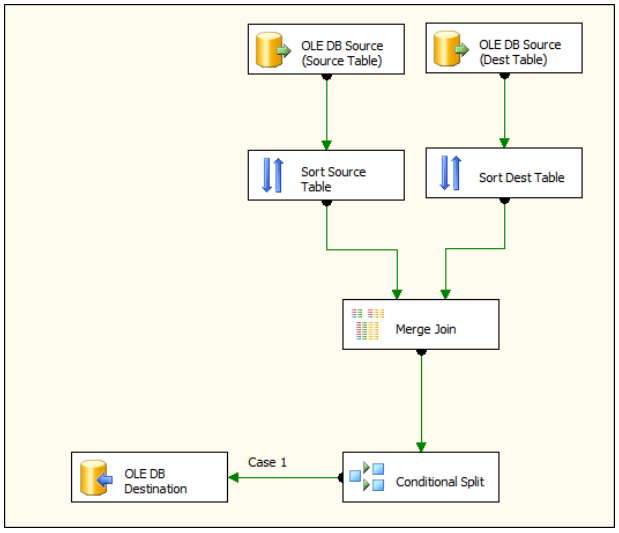


Figura 12 – Exemplo utilizando *Microsoft* SSIS IDE (Microsoft 2010).

Esta ferramenta permite criar todo o processo de carregamento, extração e transformação através de grafos, de forma simples e dinâmica, em que cada um dos seus nós é composto por um componente para efetuar uma entre as várias operações aos dados disponíveis naquele fluxo de dados. No exemplo apresentado anteriormente, foram utilizados componentes primeiramente para extrair os dados de um Sistema de Gestão de Base de Dados (SGBD) e de um ficheiro, para de seguida juntar as informações das duas fontes com o objetivo de inserir toda essa informação num determinado armazém. Ao executar o pacote, é possível facilmente identificar a quantidade de registos que passa por cada caminho do grafo, bem como compreender que todo o processo ocorreu como esperado. Para desenvolver no SSIS é necessário instalar a ferramenta MSSDT (*Microsoft SQL Server Data Tools*) para criar as várias camadas que constituem a solução de BI. É necessário ter instalado o *SQL Server Management Studio* para gerir e/ou executar num ambiente de produção (Microsoft 2015).

Por fim, também é possível realizar as seguintes operações via MSSDT (Microsoft 2015):

- Importar e exportar pacotes via assistente para criar cópia de dados de uma fonte para um destino;
- Criar pacotes com um fluxo de controlo complexo, fluxo de dados, lógica orientada e registo de eventos;
- Realizar testes e *debug* sobre os diferentes pacotes existentes, bem como monitorizar os diferentes recursos que o SSIS utiliza;
- Criar configurações que permitam alterar as definições/propriedades dos pacotes em tempo de execução;
- Instalar os pacotes e as suas dependências noutras máquinas através de um utilitário existente para o efeito;

- Gravar cópias dos pacotes numa base de dados específica, do sistema do SSIS e do sistema de ficheiros.

2.8.3 Pentaho Data Integration

O *Pentaho Data Integration* é uma ferramenta de ETL com uma abordagem orientada por meta dados. Como referido nas ferramentas anteriores, esta também não foge à regra e possui uma interface gráfica intuitiva, com componentes *drag and drop* como é possível verificar na Figura 13, seguindo uma arquitetura padrão de criação de um grafo de forma a definir um fluxo de dados (Pentaho 2015).

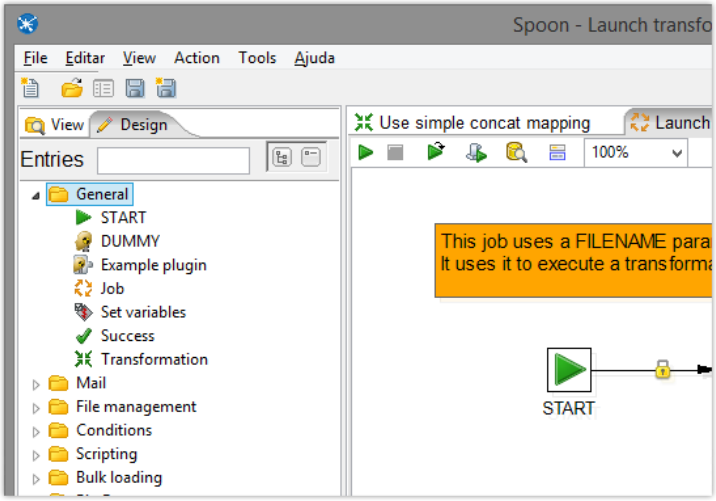


Figura 13 – Pentaho IDE.

Por fim, possui funcionalidades de monitorização, criação de perfil de dados, mecanismo de balanceamento de carga, mecanismo para reverter alterações efetuadas por *jobs* e obter dados de múltiplas fontes de dados. Esta ferramenta possui uma loja de *plugins* que se pode adicionar à solução base, permitindo assim complementar a aplicação fornecida pela *Pentaho* e integrar ou transformar dados de diferentes fontes. Não obstante, esta ferramenta possui também capacidade de elaborar soluções recorrendo ao *Big Data* (Pentaho 2015).

2.8.4 IBM – InfoSphere Information Server

O *InfoSphere* da IBM destina-se a soluções de BI, sendo um software otimizado que permite não só a integração com várias fontes de dados mas também permite realizar análises com um elevado nível de detalhe. Adicionalmente, permite também proteger os dados recorrendo a metodologias de segurança.

O desenvolvimento de soluções usando esta ferramenta é bastante intuitivo, na medida em que, possui uma interface gráfica com componentes *drag and drop* que facilitam o desenho de

todo o processo de ETL, como se pode verificar na Figura 14. Não obstante, permite integração com a *cloud*, graças ao *IBM dashDB*, que consiste em gerir todos os dados armazenando-os na *cloud* permitindo o acesso instantâneo à informação, não sendo necessário recorrer a uma ferramenta de análise para visualizar os dados (IBM 2015). Na Figura 14 é apresentado um exemplo de carregamento de dados para o AD, a partir de fontes de dados no formato XML (*Extensible Markup Language*).

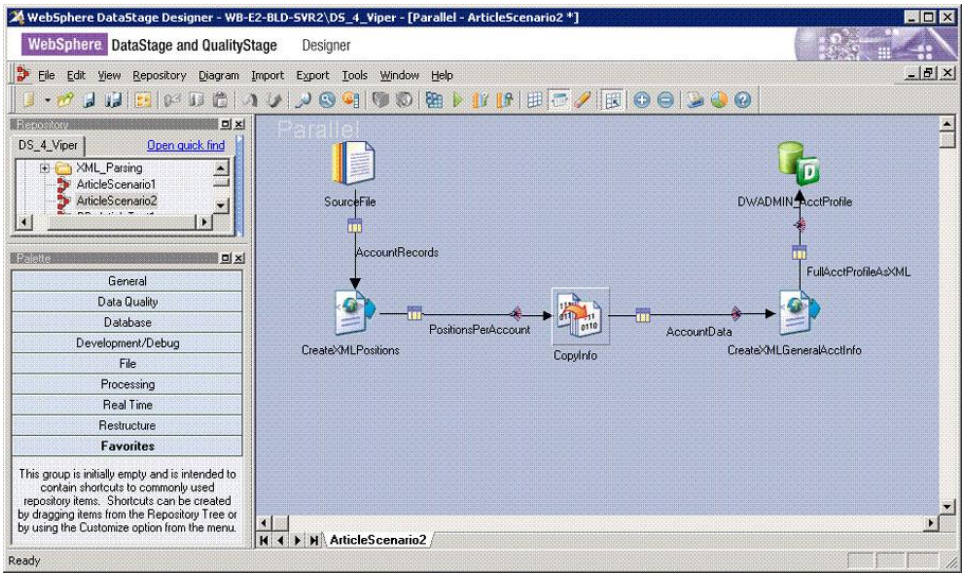


Figura 14 – IBM InfoSphere DataStage (Branislav Barnak et al. 2009).

2.8.5 Comparação de ferramentas ETL

Na Tabela 3, encontra-se a comparação das ferramentas de ETL apresentadas anteriormente. Esta tabela apresenta os prós e contras da utilização de cada ferramenta.

Tabela 3 - Comparação de ferramentas ETL ('Etl tools comparison' 2016; McBurney 2007; InformationWeek 2015).

	Vantagens	**Desvantagens**
Microsoft Integration Services (SSIS)	Vasta Documentação e Suporte; Boa performance perante grandes volumes de dados; Possui interação gráfica a nível de criação nos mapeamentos dos dados.	Compatibilidade em ambientes não Windows; Manuseamento da ferramenta pode tornar-se complexo;
Oracle Data Integrator (ODI)	Melhor conexão com aplicações de AD da Oracle; Complexidade de utilização média; Possibilidade de integração de todas as ferramentas num só ambiente aplicacional.	Complexidade de utilização; Pouca documentação; Esta ferramenta apenas se foca no processo ETL; Apenas é utilizada em *back-end* a nível de tratamento e transformação de dados, sem qualquer preocupação na disponibilização dos dados.
Pentaho Data Integration	Bom suporte fornecido pela Pentaho; Fácil utilização; Permite integração com outros produtos (BI, EII, e EAI); Performance razoável perante grandes volumes de dados; Vários objetivos de transformação e suporte aumentando as *slow changing dimensions*; Permite integração com serviços *web*.	Permite transformar os dados num *cluster*, mas não executa eficientemente o reparticionamento dos dados em paralelo; Não possui componente adequado para realizar testes de qualidade; Só consegue lidar com pequenos volumes de dados quando efetua *lookup's*; Baixo desempenho quando necessita de executar vários *scripts* de SQL em *lookup's*.
IBM – InfoSphere Information Services	Performance razoável perante grandes volumes de dados; Permite integração com serviços *web;* Visão mais forte e flexível do mercado.	Processamento em paralelo; Aprendizagem pode tornar-se complexa e difícil; Necessita de muito bom equipamento para que o processamento seja razoável.

É possível verificar que cada uma das ferramentas apresenta mais-valias no contexto de desenvolvimento de um processo de tratamento e carregamento de dados. A ferramenta da *Microsoft* possui capacidades real-time que não são muito evidenciadas pelo ODI, no entanto a ferramenta da Oracle possui um âmbito mais focado no ETL e processamento paralelo. No caso da ferramenta *Pentaho* e IBM, apesar de serem ferramentas com um bom suporte fornecido por parte da empresa que o desenvolve, estes contêm problemas de performance quando perante elevado volume de dados.

2.9 Ferramentas de apresentação/analíticas

Depois de extraídos e tratados os dados das fontes para o novo sistema de gestão de informação, podem ser utilizadas algumas ferramentas com o objetivo de auxiliar na sua análise. Estas ferramentas existem no mercado para responder às necessidades de análises de grandes quantidades de dados, com muito pouco detalhe técnico e sobretudo com capacidade abstrata, adaptando-se a diferentes realidades. De seguida é apresentado um estudo feito de algumas das ferramentas mais utilizadas pelo mercado.

2.9.1 Microstrategy

A *MicroStrategy Analytics*™ é uma ferramenta analítica comercializada pela *MicroStrategy* que permite a análise e a partilha de conhecimento a partir de grandes volumes de dados. Possui uma interface familiar ao utilizador comum, permitindo de forma simples e prática a geração de relatórios e *dashboards,* apresentando e medindo a informação pretendida. Atualmente, é bastante utilizada por várias organizações a nível mundial, como o Ebay, a Zurich, o Novo Banco, a Portugal Telecom, entre outras organizações (MicroStrategy 2015).

Numa primeira fase, são apresentados um conjunto de opções com o objetivo de definir quais as fontes de dados alvo. Tomando como exemplo uma importação de dados de um ficheiro do tipo Excel e depois de escolhidas as folhas onde é necessário extrair informação, a ferramenta retira os atributos existentes e determina possíveis métricas a serem medidas pelo tipo de estrutura apresentado. É possível alterar a estrutura pré-deduzida.

Numa fase seguinte, o utilizador é livre de definir quais os atributos e métricas a analisar. Possui uma interface apelativa e fácil de utilizar, na medida em que os atributos e métricas são selecionados para o *dashboard* através de *drag and drop.* Cada um destes pode ser personalizado com vários tipos de gráficos, por forma a embelezar e tornar percetível o que se está a analisar, para posterior avaliação e tomada de decisão. Também, contém métodos internos que realizam algumas transformações aos dados, por forma a realizar as análises com base na estrutura de informação definida, nomeadamente análises preditivas.

Na Figura 15 é apresentado um exemplo de utilização, disponibilizado pela própria empresa, relacionado com o ciclismo. Apresenta a média de utilizadores por hora do dia, dia da semana, por mês e ainda a comparação do número de utilizadores versus a temperatura do dia.

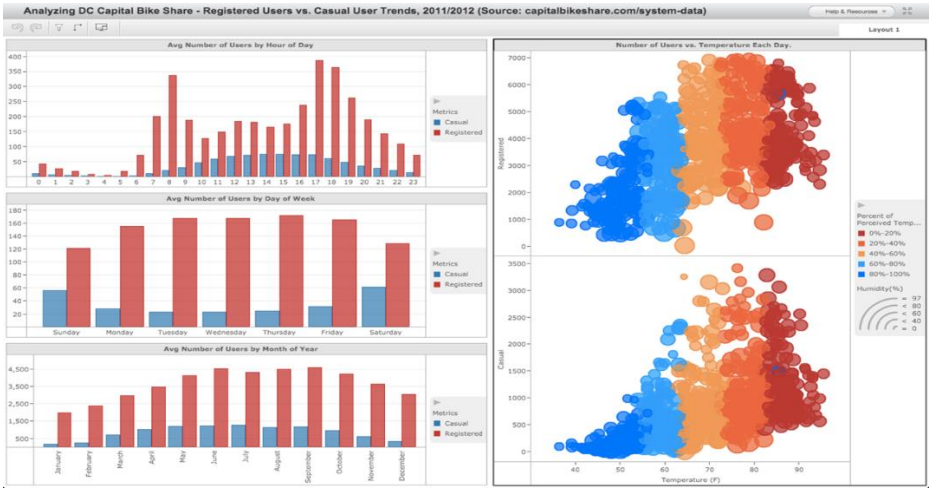


Figura 15 – Exemplo de utilização *MicroStrategy Analytics™ desktop* (MicroStrategy 2015) .

Para além da aplicação *desktop*, também permite que toda esta informação seja acedida por dispositivos móveis, facilitando assim essencialmente a partilha de informação entre grupos de trabalho.

Por fim, é de realçar que os preços de licenças desta ferramenta são elevados, não permitindo que qualquer empresa possa usufruir destas funcionalidades. Também, como limitação, não existe aplicação móvel para *Windows Phone,* o que a impede ser uma solução *cross-platform.*

2.9.2 Microsoft Power BI

Esta é outra das soluções analíticas encontradas, simples de usar e gratuita. O *Microsoft Power BI* é um conjunto de funcionalidades disponibilizadas online ou localmente em aplicação *desktop*, que permite a criação de *dashboards* interativos através de conjuntos de informação, independentemente do tipo e do tamanho. As criações feitas podem ser partilhadas entre utilizadores e publicadas para que qualquer pessoa consiga usufruir (Microsoft 2016a). Numa primeira abordagem, o utilizador começa por definir qual a fonte de dados. Possui uma lista variada de tipos de fontes que são aceites, permitindo ainda combinar dados entre fontes diferentes.

Na Figura 16 é apresentado um exemplo de utilização da ferramenta, no contexto da gestão do desenvolvimento feita a partir do *Visual Studio* (*framework* de desenvolvimento).

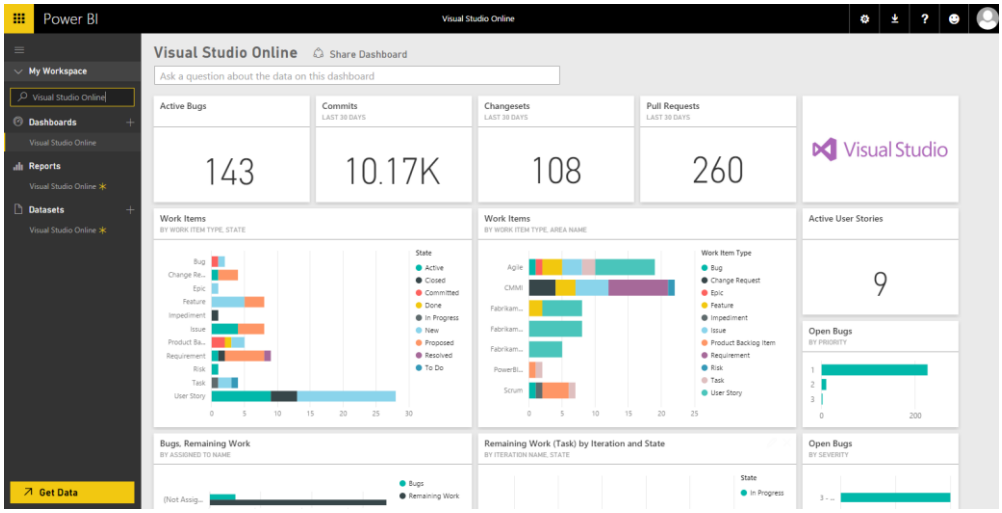


Figura 16 – Exemplo de utilização *Microsoft Power BI*.

2.9.3 Phocas

Como alternativa à solução apresentada anteriormente, o *software* Phocas disponibiliza funcionalidades capazes de efetuar análises e partilhar resultados. É composto por um módulo de análise de dados, onde o utilizador seleciona o tipo de informação a importar, permitindo realizar todo o processo de análise na nuvem de dados e a partir de aí mostrar os dados aos utilizadores finais em diferentes plataformas personalizáveis. Apresenta uma interface simples e prática, apta a ser utilizada mesmo por utilizadores sem qualquer tipo de formação técnica específica (Phocas 2015).

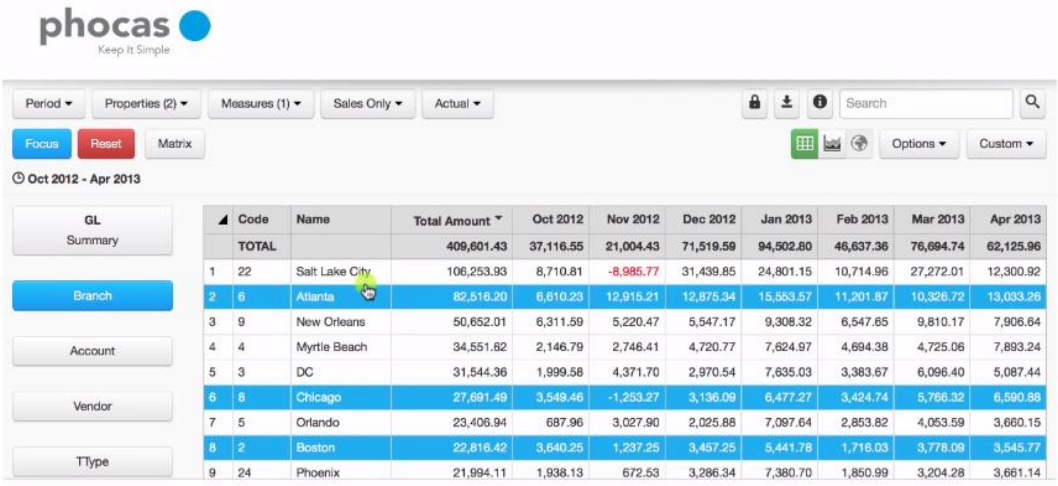


Figura 17 - Overview do componente de análise (Phocas 2015).

Como é demostrado na Figura 17, este componente determina o tipo de medidas que podem ser usadas nas análises, o tipo de filtros de informação que podem ser criados e permite a escolha da forma como as análises são apresentadas, nomeadamente em forma de grelha,

tabelas, gráficos e mapas (no caso em que seja possível extrair informação agregada de localização sobre os dados).

A partilha do conhecimento dessas análises pode ser feita através do módulo para dispositivos móveis, definindo para que utilizadores a informação estará disponível para interação ou apenas visualização. Esta partilha também pode ser composta por *dashboards* e relatórios mais elaborados, criados a partir de uma funcionalidade disponível pela aplicação. A partir da fonte de dados anteriormente definida, filtrada e com todas as medições pretendidas, o utilizador pode fazer *drag-and-drop* de diversos componentes, definindo para cada um deles o tipo de informação a disponibilizar (Figura 18).

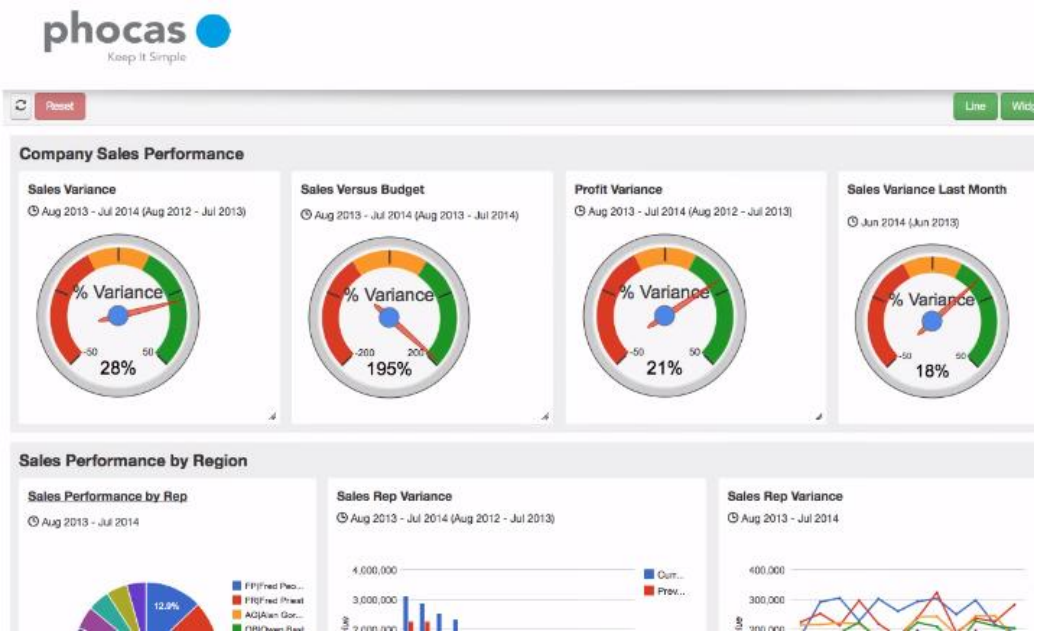


Figura 18 - *Overview* da criação de dashboards (Phocas 2015).

Esta solução tem como vantagens o facto de ser escalável em qualquer altura, permitindo a alteração das estruturas e do tipo de informação das análises de forma bastante facilitada e facilmente acessível em qualquer dispositivo, através de um portal *web*. Permite ainda a integração com sistemas ERP (*Enterprise Resource Planning)* e CRM (*Customer Relationship Management*) (Phocas 2015).

Não obstante, disponibiliza uma camada de segurança com autenticação via LDAP (*Lightweight Directory Access Protocol*), com camada de administração para gestão de acessos e ainda segurança a nível de *deploy* para a *cloud*, com recurso a *Rackspace (*tecnologia que segue várias normas de segurança ISSO (*International Organization for Standardization*), sendo por isso uma mais-valia)(Nubera 2015).

Por fim, como desvantagem este necessita de alojamento em servidores Windows como requisito mínimo, não sendo por isso uma solução *cross-platform* neste aspeto. Os custos de licenças e alocação da solução são aspetos também menos vantajosos, por mais uma vez a

alocação da solução apenas poder ser feita em servidores *Windows*, sendo necessário despender mais recursos monetários para estes (Nubera 2015).

2.9.4 Comparação entre ferramentas analíticas de apresentação de dados

Na Tabela 4 encontra-se a comparação entre as diferentes ferramentas analíticas existentes no mercado. Esta comparação segue os seguintes parâmetros, que são considerados importantes para uma ferramenta deste tipo (G2 Crowd 2016):

- **Usabilidade:** facilidade com que se criam novas análises;
- **Configuração:** configuração da ferramenta na máquina onde se realizam as análises;
- **Manutenção:** facilidade com que se mantém as análises existentes;
- **Suporte:** suporte fornecido à ferramenta por parte do fabricante;
- **Fácil para o negócio:** facilidade de utilizar para o negócio;
- ***Self-Service*:** ser o mais fácil de ser utilizado pelo utilizador;
- **Análises avançadas:** possibilidade de efetuar análises complexas;
- **Construção de novos relatórios:** facilidade de construção de novos relatórios.

Tabela 4 - Comparação de ferramentas analíticas (G2 Crowd 2016).

	Microstrategy	**Phocas**	**Power BI**
Usabilidade	Complexo	Intuitivo e Fácil	Intuitivo e Fácil
Configuração	Difícil de configurar	Simples de configurar	Simples de configurar
Manutenção	Difícil de manter	Fácil de Manter	Fácil de Manter
Suporte	Contém um bom suporte	Suporte muito bom	Suporte Reduzido
Fácil para negócio	Contém alguma complexidade	Muito fácil	Fácil
Self-Service	Não é uma ferramenta fácil de utilizar	Muito Intuitiva	Intuitiva
Análises Avançadas	Suporta análises complexas	Bom suporte em análises complexas	Não lida muito bem com análises complexas
Construção de relatórios	Um pouco complexo	Bastante fácil e intuitivo	Fácil e intuitivo
Mais utilizado	Empresas	Pequenas- Médias Empresas	Pequenas- Médias Empresas

Desta forma, consegue-se apurar que a ferramenta Phocas é mais completa e simples, desde a interação por parte do utilizador até à manutenção da informação utilizada nas análises de dados criadas.

2.10 Análise de Mercado

Esta secção tem como objetivo apresentar possíveis soluções ou estudos que foram desenvolvidos e postos em prática no âmbito da criação e manutenção de um AD no contexto escolar. Numa segunda abordagem são apresentadas aplicações de mercado, sendo vistas como uma possível solução para apresentação dos dados armazenados. Destacam-se pela sua baixa complexidade de compreensão/uso, bem como pela completa gama de funcionalidades que disponibilizam, de acordo com as necessidades do meio em que são utilizadas.

Para além destes, existem outros indicadores que são específicos das instituições, como é o caso da LEI. Esse conjunto de indicadores serão apresentados numa próxima secção.

2.10.1 Ferramentas Comerciais de Gestão para Ensino

Nesta secção descrevem-se alguns dos produtos existentes atualmente no mercado para montagem/criação de um armazém de dados.

2.10.1.1 EdVantage

Desenvolvido pela *VersiFit Technologies*, este é um produto que tem a gestão de dados curriculares como objetivo principal. É possível avaliar todo o tipo de informação referente aos alunos, aos professores e a cada programa de disciplina que foi efetivamente lecionado. Para além da disponibilização do armazém de dados para armazenamento da informação, este produto permite ainda o uso de um conjunto de funcionalidades, tais como: gerar *reports* de acordo com o tipo de informação que o utilizador pretende consumir, definir métricas comparativas, *drill-down* de análises, interrogar as consultas através de uma ferramenta avançada incluída no produto e adaptar o produto às necessidades do consumidor (SchoolCity Inc. 2015). A Figura 19 apresenta a estrutura geral deste produto, adaptado a qualquer contexto. Foi encontrada a utilização deste produto por parte da Escola Elementar de *Buffalo* (School Buffalo 2015).

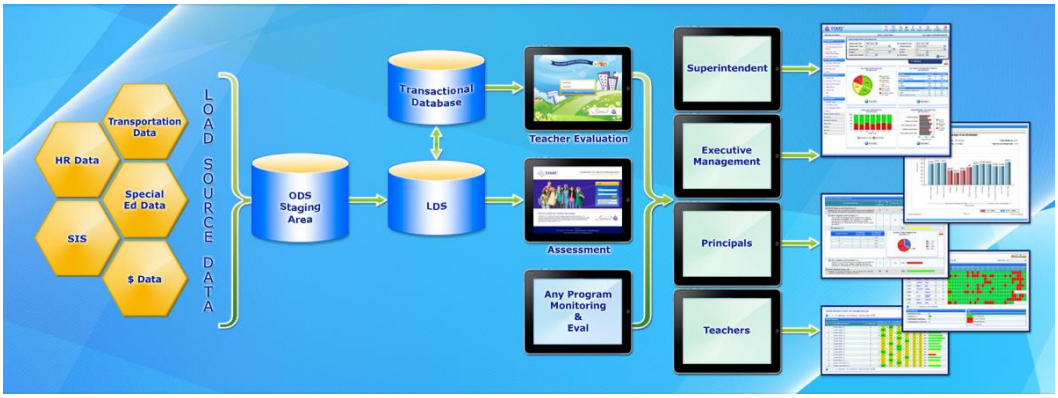


Figura 19 – EdVantage (SchoolCity Inc. 2015) .

2.10.1.2 DataCation

DataCation (CaseNex 2010) é um AD desenvolvido pela CaseNex. Permite reunir dados num local centralizado, dados estes no contexto escolar e que normalmente se encontram separados fisicamente (informação dos alunos, dos Recursos Humanos, Finanças). Os dados são carregados no sistema, tendo como origem os sistemas atualmente existentes na instituição e utiliza o processo ETL para esse efeito. Desta forma consegue-se garantir a integridade e confiança dos dados (CaseNex 2010).

Abstraindo do sistema, a empresa disponibiliza um conjunto de ferramentas que se baseia nos dados armazenados que serão apresentados de forma agregada. Dependendo do tipo de utilizador final que cada ferramenta vai ter, foram consideradas métricas *standard* essenciais que servem como base para as tomadas de decisões e alinhamento dos objetivos/alterações necessárias na instituição. No caso em que sejam necessários novos parâmetros de análise, esta também disponibiliza uma solução em específico, tornando esta solução escalável e adaptável às necessidades (CaseNex 2010).

Em suma, destacam-se as seguintes ferramentas/módulos disponibilizadas por parte da empresa:

- ***Skedula - School/Teacher Management Portal***: Ferramenta utilizada por parte dos docentes permitindo seguir o percurso do aluno a nível de avaliações, trabalhos de casa, evoluções e eventos que tenha associado (Figura 20);

- ***Block sckedula - SIS***: Orientada para os docentes administrativos, permitindo criar o processo de admissão do aluno, consultar e fazer a manutenção do registo (disciplinas a que está inscrito, faltas, registos de saúde, entre outros itens). Permite a criação de relatórios estatísticos sobre qualquer informação do aluno;

- ***Graduation Eligibility Tracking System*:** Permite fazer a gestão da informação de resultados de avaliações a nível do aluno e das escolas. A partir desta informação é

possível analisar a performance de uma determinada escola de um agrupamento por tipo de resultado, por etnia, por género, entre outros parâmetros;

- ***Pupil Path-Parent/Student Portal*:** Plataforma *online* que permite aceder à informação do aluno por parte dos pais ou mesmo pelo próprio aluno. Permite consultar as avaliações, os programas que o aluno frequenta, objetivos de aprendizagem, entre outros. Esta informação pode também ser extraída em formato de relatório, sendo este designado como um relatório de progresso do aluno;

- ***Scorecard/Dashboard*:** Permite a criação de novas métricas associadas ao agrupamento/escolas, com o objetivo de detetar novos pontos de interesse e assim melhorar os serviços oferecidos pelas instituições de ensino. Permite analisar em tempo real as novas métricas definidas e analisar os KPI entre estudantes, agrupamentos e escolas;

- ***K-12 Diagnostic Management System*:** Orientada para os administradores de cada agrupamento, permite auxiliar os professores nas análises que efetuam relativamente a avaliações medíocres/elevadas, identificando logo à partida os estudantes em risco (medindo o seu desempenho numa unidade curricular em particular);

- ***PADS-Grade Management System***: Solução *online* que permite gerir todos os relatórios possíveis de serem gerados;

- ***Assessment Management System*:** Sistema responsável por recolher todos os dados possíveis sobre o desempenho dos alunos nas entregas ao longo do ano curricular;

- ***Teacher Performance Record*:** Sistema responsável por recolher todos os dados possíveis sobre o desempenho dos docentes em sala de aula.

Figura 20 – Exemplo de utilização da aplicação *Skedula - School / Teacher Management Portal*(CaseNex 2010).

2.10.2 AD na área do ensino/educação

Nesta secção são apresentadas algumas soluções utilizadas em contexto de ambientes escolares.

2.10.2.1 University Data Warehouse Plus

A Universidade de Nova Iorque desenhou um sistema de apoio à decisão, interno à instituição. Este sistema gere diferentes tipos de informação vindos de departamentos/áreas diferentes e, para isso é disponibilizado aos utilizadores finais um conjunto de ferramentas analíticas e de *reporting*, capazes de responder às necessidades operacionais (New York University 2014). Possui quatro áreas de destaque:

- **Informação de Finanças**: este módulo foi o primeiro a avançar no que toca ao desenvolvimento e engloba a criação de relatórios e *dashboards* de informação agregada financeira. Tem como utilizadores alvo presidentes, chefes de departamento, investigadores, administradores, entre outros. Para além de usufruírem do sistema, podem ser atores importantes na criação/desenvolvimento do sistema, na medida em que podem definir as necessidades que precisam de ser satisfeitas, lista de KPI que são importantes de medir e que tipo de informação é importante reter. Numa fase posterior é criado o novo relatório ou componentes que podem ser usufruídos por qualquer um dos utilizadores que pertença a esta área;

- **Métricas por Departamento da Instituição**: apresenta informação sobre professores, investigadores e alunos, aplicando métricas definidas pelo *Provos't Council on Science and Technology*, orientado para chefes de departamento, reitores e outras instituições;

- **Informação de Recursos Humanos**: utiliza informação sobre os docentes, as posições atribuídas a cada docente na instituição, a distribuição de trabalho e os salários dos colaboradores;

- **Informação referente aos alunos:** O módulo corrente é integrado com os anteriores, focando-se apenas sobre os dados dos alunos e respetivo historial curricular.

Atualmente encontram-se em fase de redesenho do corrente AD, com o objetivo de criar novas estruturas de dados para armazenar novos tipos de informação, de acordo com as necessidades que foram sendo encontradas. Estas alterações têm como objetivo tornar o sistema mais simples de usar bem como mais íntegro, seguro e rápido de utilizar em situações emergentes. Outra das alterações efetuadas encontra-se ao nível da ferramenta de *reporting* utilizada. De forma a capacitar o sistema de uma gama mais completa de relatórios e capacidades analíticas diversificadas, passaram a utilizar a ferramenta da *Oracle*, mais designadamente a *Oracle Business Intelligence Enterprise Edition* (New York University 2014). Na Figura 21 apresenta-se uma imagem da arquitetura criada.

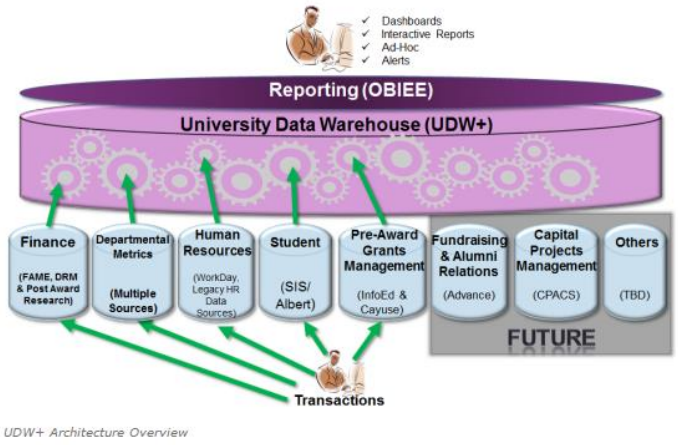


Figura 21 - Arquitetura do AD da Universidade de Nova Iorque (New York University 2014).

2.10.2.2 School District of Philadelphia

O agrupamento de escolas do distrito de Filadélfia necessitava de captar e armazenar os dados curriculares de uma forma simples e centralizada, com o objetivo de melhorar todo o processo escolar, apoiando os docentes nas suas operações diárias. Devido ao número de escolas existentes no agrupamento (encontra-se divido em onze regiões / duzentas e setenta escolas no total), a gestão e o acesso à informação por parte de todo o ambiente docente é difícil, sendo considerada uma das sete maiores escolas dos Estados Unidos. Desta forma, a IBM uniu-se com a Universidade de Pensilvânia e com a própria escola com o objetivo de criar um AD que respondesse às suas necessidades, criando atualmente relatórios e *dashboards* mensais/anuais, que otimizam o processo e resumem cerca de setecentos indicadores identificados como essenciais (IBM 2006).

Este novo sistema trouxe com ele melhorias no ambiente escolar significativas: permitiu que a taxa de absentismo dos professores diminuísse em dois por cento; a taxa de comparência nas aulas por parte dos alunos aumentou em cinquenta e sete por cento; as suspensões diminuíram em vinte e três porcento; Incidentes graves diminuíram em dezoito porcento (IBM 2006).

Para que todos os utilizadores tirem melhor partido das vantagens do novo sistema, foi criada também uma ferramenta de acesso, que gera os relatórios de acordo com as métricas estipuladas, dada pelo nome de SchoolStat (IBM 2006), orientando-se para resultados e para o desempenho da escola. Apesar da ferramenta ter a capacidade de deduzir possíveis métricas interessantes a partir dos dados carregados no sistema, foi definida uma lista de KPI base por parte do agrupamento, destacando-se os seguintes (Patusky, Botwinik & Shelley 2007):

- Percentagem de respostas corretas nos testes de *benchmark*. A elaboração dos testes por parte dos alunos é feita na disciplina de Matemática, Leitura e Ciências a cada 6 semanas, apenas destinada a alunos do ensino secundário que possuam uma classificação de seis a oito;

- Percentagem de alunos com um determinado nível atribuído pelo professor a nível da leitura;

- Número de alunos que se encontram na categoria dois (baixa aderência às aulas lecionadas, baixo desempenho ou determinado comportamento do aluno) do *Comprehensive Student Assistance Process (CSAP)* – este processo ajuda na identificação da necessidade de ensino especial para os alunos que se incluam nesta categoria;

- Taxa de ausência dos funcionários da instituição (docentes e não docentes) por dia;

- Taxa de aderência às aulas por parte dos alunos;

- Percentagem de suspensões.

Em suma, a informação recolhida de cada uma das escolas entra para o AD existente (introduzidos por parte do secretariado, pelos professores ou outros membros administrativos, no máximo mensalmente) e é a partir desta que os dados são agregados e apresentados em forma de relatórios, fáceis de interagir e compreender por parte do utilizador final (Patusky, Botwinik & Shelley 2007).

2.10.3 Indicadores utilizados no Ensino Superior

No ecossistema do ensino superior existem várias instituições de ensino com diferentes metodologias. Segundo o documento elaborado por (Saúde et al. 2014), existem várias referências de indicadores de desempenho nas instituições de ensino superior. Estes são utilizados de acordo com o(s) objetivos(s) específico(s) no que toca à fórmula para efetuar o cálculo, fontes de informação e necessidade de evidências complementares.

Neste âmbito, foi feita uma pesquisa em algumas entidades relacionadas com o ensino em Portugal, com o objetivo de apresentar um conjunto de indicadores oficiais que estas entidades consideram como fundamentais para desmistificar e controlar o acesso e o ensino. Foi possível encontrar um documento elaborado pela FenProf que apresenta o estudo relativo ao acesso ao ensino superior nos anos letivos 2009/2010 e 2010/2011. A Tabela 5 representam alguns desses indicadores.

Tabela 5 – Listagem de alguns indicadores retirados da FenProf (FenProf 2012).

Indicador	Descrição
Número de ofertas por instituição	Contagem da quantidade de cursos disponíveis por instituição (incluindo Universidades e Politécnicos).
Número de inscritos por tipo de instituição	Contagem do número de inscrições total na Universidade e no Politécnico.
Número de candidatos por fase de candidatura	Contagem do número de candidaturas recebidas em cada uma das fases.
Determinar as áreas mais procuradas por território e por distrito	Contagem do número de candidaturas de alunos com o objetivo de determinar as áreas de curso que obtiveram mais candidaturas, por área territorial.
Determinar as áreas com mais colocados	Contagem do número de alunos que submeteram as candidaturas e foram colocados, por área de curso escolhido.
Número de candidaturas recebidas por tipo de instituição	Contagem do número de candidaturas recebidas no total nas Universidades e nos Politécnicos.
Número de alunos inscritos por tipo de instituição	Número de alunos inscritos no total, mas Universidades e Politécnicos.
Contagem da oferta disponível por instituição	Número de cursos que a universidade/politécnico oferece aos alunos (atualmente em curso, descontinuados e à espera de aprovação) em Licenciatura e Mestrado, nas instituições públicas e privadas.
Número de vagas por tipo de instituição	Contagem do número de vagas disponíveis total, em Universidades e Politécnicos.

3 Análise e Desenho

Neste capítulo será apresentada toda a abordagem técnica efetuada sobre o objetivo da dissertação. Na primeira secção, será descrito de forma breve o problema a resolver e a respetiva análise, identificando o formato das fontes de dados e que alternativas existem para contornar os problemas na extração dos dados necessários.

Na segunda secção serão apresentadas as fontes de dados disponibilizadas pelo cliente e os principais indicadores a analisar, extraídos dos exemplos de análises igualmente fornecidos.

No bloco seguinte, será definida a arquitetura inicial, enquadrando cada uma das camadas especificadas para a solução. De notar que nem todas as estruturas de dados apresentadas vão ser criadas neste contexto específico de avaliações, devido ao facto de existir um outro *data mart* desenvolvido em paralelo por um outro colega. Apenas será feita uma breve descrição, tendo em conta que os dados armazenados vão ser relacionados com os artefactos apresentados.

3.1 Descrição do problema

Atualmente, como foi descrito no capítulo inicial, o armazenamento e análise de dados é feito a partir de um ficheiro Excel, extraído com uma estrutura de dados a partir do Portal da instituição. Dado o crescimento exponencial de dados, a manutenção de análises aos mesmos através deste tipo de ficheiros torna-se difícil de gerir. Para cada análise que se pretenda estudar é necessária a aplicação de fórmulas matemáticas e respetivos filtros, aumentando o problema quando existe a possibilidade de acrescentar novos campos de dados, não só na possível aplicação de regras de cálculos sobre esses novos dados, mas também porque exige uma nova exportação. Da mesma forma a pesquisa de padrões e comportamentos torna-se

arriscada, de perceção lenta devido ao número elevado de registos que pode conter, forjada na medida em que não existem garantias de limpeza, uniformização, tratamento dos dados e de erros que possam ser criados a partir dos cálculos, corrompendo a integridade e autenticidade dos dados.

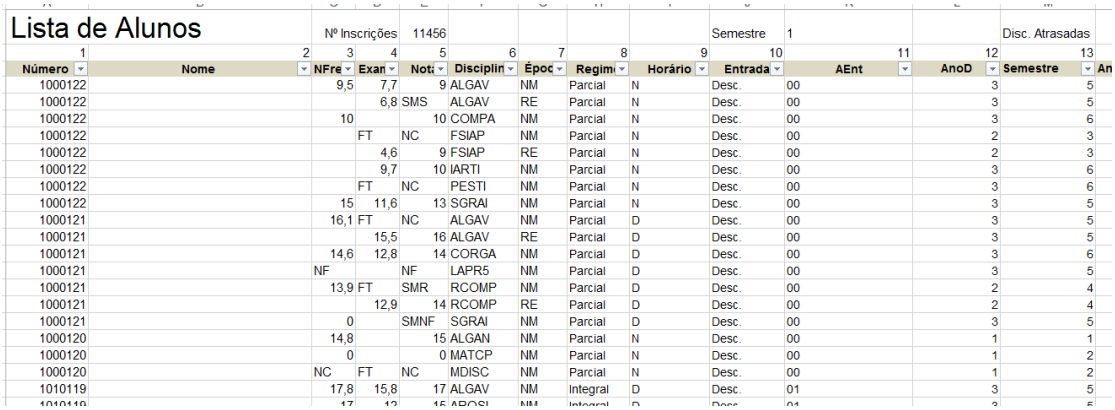
Assim, surgiu a necessidade de criar um sistema que permita concorrentemente a confidencialidade dos dados (na medida em que fica à responsabilidade do Diretor de curso), o apoio nas tomadas de decisões, bem como na apresentação de resultados, escalabilidade e a fácil manutenção de todo um volume de dados. Na Figura 22 é apresentado de uma forma breve os intervenientes e as interações possíveis com a solução pretendida.

Figura 22 – O processo de extração, intervenientes e interações com a solução pretendida.

3.2 Fontes de Informação

Os suportes de dados disponibilizados para o carregamento de informação são constituídos unicamente por folhas de cálculo. Nelas existe um conjunto de informação sobre a LEI que é exportado a partir do Portal da instituição, e como tal nem todo o conteúdo é relacionado com o que é pretendido armazenar. Para isso, neste contexto específico e de acordo com as orientações dadas pelo cliente (Diretor de curso da LEI), apenas foram consideradas as seguintes folhas de informação: página de informação de avaliação de alunos, página com informação de todas as unidades curriculares e, por fim, uma página que contém a informação de inscrição dos alunos (apenas com classificação final). Foi disponibilizada informação em dois intervalos de tempo distintos: ano letivo 2012/2013 e 2013/2014.

A título de exemplo, a primeira folha de cálculo mencionada (Figura 23) contém a informação de todas as avaliações dos alunos no ano letivo 2013-2014. É possível observar com que notas foram concluídas as várias disciplinas a que estava inscrito, de acordo com os tipos de avaliação da disciplina (época de avaliação e classificação atribuída), o ano corrente e o de conclusão, o horário, entre outras informações.



Lista de Alunos — Nº Inscrições 11456 — Semestre 1 — Disc. Atrasadas

Número	Nome	NFre	Exan	Noti	Disciplin	Époc	Regim	Horário	Entrada	AEnt	AnoD	Semestre	An
1000122		9,5	7,7	9	ALGAV	NM	Parcial	N	Desc.	00	3	5	
1000122			6,8	SMS	ALGAV	RE	Parcial	N	Desc.	00	3	5	
1000122		10		10	COMPA	NM	Parcial	N	Desc.	00	3	6	
1000122			FT	NC	FSIAP	NM	Parcial	N	Desc.	00	2	3	
1000122			4,6	9	FSIAP	RE	Parcial	N	Desc.	00	2	3	
1000122			9,7	10	IARTI	NM	Parcial	N	Desc.	00	3	6	
1000122			FT	NC	PESTI	NM	Parcial	N	Desc.	00	3	6	
1000122		15	11,6	13	SGRAI	NM	Parcial	N	Desc.	00	3	5	
1000121		16,1	FT	NC	ALGAV	NM	Parcial	D	Desc.	00	3	5	
1000121			15,5	16	ALGAV	RE	Parcial	D	Desc.	00	3	5	
1000121		14,6	12,8	14	CORGA	NM	Parcial	D	Desc.	00	3	6	
1000121		NF		NF	LAPR5	NM	Parcial	D	Desc.	00	3	5	
1000121		13,9	FT	SMR	RCOMP	NM	Parcial	D	Desc.	00	2	4	
1000121			12,9	14	RCOMP	RE	Parcial	D	Desc.	00	2	4	
1000121		0		SMNF	SGRAI	NM	Parcial	D	Desc.	00	3	5	
1000120		14,8		15	ALGAN	NM	Parcial	N	Desc.	00	1	1	
1000120		0		0	MATCP	NM	Parcial	N	Desc.	00	1	2	
1000120		NC	FT	NC	MDISC	NM	Parcial	N	Desc.	00	1	2	
1010119		17,8	15,8	17	ALGAV	NM	Integral	D	Desc.	01	3	5	

Figura 23 - Folha de cálculo com informação detalhada de avaliação.

Na Figura 24 é apresentada a folha designada de "configurações" de disciplinas e tem como objetivo determinar quantos ECTS possui cada unidade curricular, em que ano curricular é lecionada e em que semestre.

Disciplinas	Ano	FAA	Semestre	NM	RE	EE	TE	EF	Acrescentar	ID	ECTS
ALGAN	1		1							1	5
APROG	1		1							2	5
AMATA	1		1							3	5
LAPR1	1		1							4	8
PRCMP	1		1							5	6
ARQCP	2		3							6	5
BDDAD	2		3							7	6
ESINF	2		3							8	6
FSIAP	2		3							9	5
LAPR3	2		3							10	8

Figura 24 - Folha de cálculo com informação sobre as disciplinas.

Por último, resta a fonte de dados que incide sobre as inscrições dos alunos (Figura 25, apenas com registo de classificação final), contendo dados sobre o aluno numa inscrição a determinada cadeira, a data de finalização da mesma, a classificação obtida e a informação sobre creditação de competências (se existir). De notar que esta informação não é do mesmo intervalo de tempo que as fontes indicadas acima. Até aqui a informação disponibilizada é referente a um ano letivo. Neste caso, podem conter datas de finalização fora deste intervalo que, perante o sistema, apesar de não reunir toda a informação necessária, vão ser consideradas como uma inscrição e uma avaliação dado que possui uma nota final associada.

Número	Disciplina	No	Data	TN	AnoEnt	AnoDisciplina	AuxNumUC	ECTS	NotaPesad	AuxTotNoti	CreditosAu
1000122	ALGAN	12	22/01/2009		00	1	1	5	60	60	5
1000122	ALGAV	10	02/09/2014		00	3	2	5	50	110	10
1000122	AMATA	11	10/02/2010		00	1	3	5	55	165	15
1000122	APROG	13	13/02/2007		00	1	4	6	78	243	21
1000122	ARQCP	11	09/02/2011		00	2	5	5	55	298	26
1000122	ARQSI	12	15/02/2013		00	3	6	5	60	358	31
1000122	ASIST	10	07/02/2013		00	3	7	5	50	408	36
1000122	BDDAD	12	08/02/2012		00	2	8	6	72	480	42
1000122	COMPA	10	12/04/2014		00	3	9	4	40	520	46
1000122	CORGA	12	02/07/2013		00	3	10	4	48	568	50
1000122	EAPLI	13	11/07/2012		00	2	11	6	78	646	56
1000122	ESINF	13	07/12/2012		00	2	12	6	78	724	62
1000122	ESOFT	12	29/06/2010		00	1	13	6	72	796	68
1000122	GESTA	12	28/01/2013		00	3	14	4	48	844	72
1000122	IARTI	10	04/07/2014		00	3	15	4	40	884	76

Figura 25 - Folha de cálculo com a informação de inscrição dos alunos (apenas com classificação final).

3.3 Análises

No âmbito deste projeto, para além das fontes de dados necessárias a serem carregadas no sistema, também foi disponibilizada informação pelo cliente sobre o tipo de análises mais importantes e mais frequentes no contexto de avaliações de alunos que decorrem com alguma periodicidade. Para além das disponibilizadas foram criadas novas análises com o objetivo de melhor comprovar mais para a frente a veracidade do sistema. É de salientar que todas elas são filtradas de acordo com o ano letivo que se pretende analisar e sempre tendo em conta a LEI.

Deste conjunto de análises, foram então consideradas as seguintes:

A. **Unidades curriculares inscritas versus unidades aprovadas:** contagem de alunos com um determinado número de inscrições e com um número de unidades curriculares aprovadas;

B. **Unidades curriculares inscritas versus unidades reprovadas**: contagem de alunos com um determinado número de inscrições e com um número de unidades curriculares reprovadas;

C. **Unidades aprovadas por ano curricular:** número de alunos que obtiveram um número de aprovações em unidades curriculares, por ano curricular;

D. **Unidades curriculares em atraso por ano curricular:** contagem do número de alunos que tem um determinado número de disciplinas em atraso por ano curricular;

E. **Aprovações por ano curricular e regime:** consiste no número de alunos que obtiveram um número de aprovações em disciplinas, por ano curricular e regime ao qual frequentou;

F. **Aprovações por ano curricular e horário:** consiste no número de alunos que obtiveram um número de aprovações em disciplinas, por ano curricular e horário ao qual frequentou;

G. **Contagem de avaliações por tipo de classificação obtida, por disciplina e por semestre:** Informação de classificações detalhadas de um ano letivo. Indica o número de alunos que obtiveram, por semestre e por cada unidade curricular, uma de entre as seguintes classificações: “Aprovado”, “Reprovado”, “SMR” (Sem nota mínima recurso), “SMS” (Sem nota mínima Setembro), “SNMF” (Sem nota mínima e não frequentou), “NC” (Não Classificado), “NF” (Não Frequentou), “DT” (Desistiu), “FT” (Faltou), e “CC” (Creditação de Competências). Desta forma, é possível avaliar o que se passa em cada unidade curricular, verificando qual é que está a ter melhor/pior resultado. Também foi feita a análise apenas para os alunos cuja nota de frequência é positiva;

H. **Contagem de avaliações com nota frequência positiva por tipo de classificação obtida, por disciplina e por semestre:** esta análise possui o mesmo âmbito que a anterior,

contabilizando apenas registos de avaliação cuja nota de frequência seja superior a dez valores;

I. **Contagem de exame por ano letivo:** permite contabilizar o número de exames feitos por ano letivo e por tipo de exame: exame de época norma, recurso ou especial;

J. **Média de alunos finalistas:** Contagem do número de alunos finalistas que obtiveram determinada média de classificações, entre dez e vinte valores;

K. **Média de notas de alunos:** Contagem do número de alunos no geral que obtiveram determinada média de classificações, entre dez e vinte valores.

3.4 Análise da solução

Depois de apresentada toda a informação necessária, foi feita uma análise ao tipo de dados e respetivo detalhe com o objetivo de moldar a solução ao contexto de acordo com o que é pretendido. O sistema deve assim ser capaz de processar e armazenar grandes quantidades de informação sobre a avaliação dos alunos nas unidades curriculares que frequentaram, considerando cada avaliação existente parte do processo de negócio estipulado.

Cada processo de negócio diz respeito a um aluno, que frequentou a unidade curricular num determinado regime, curso (que a unidade curricular pertence), num determinado ano letivo e com determinada data do dia do lançamento da nota final. Pode-se dizer que todos estes domínios de informação vão ser transformados em dimensões.

Na avaliação, o aluno obtém uma nota de frequência que resulta do trabalho desenvolvido durante o semestre e, dependendo das regras da disciplina, é submetido a exame durante a época de exame normal. No caso em que o aluno reprove ou queira fazer melhoria de nota, resta a época de recurso. Se este estiver ao abrigo de condições especiais ainda tem acesso à época especial de exame. Resumindo, cada processo de avaliação a uma determinada disciplina é composto por uma nota de frequência, notas de exame nas diferentes épocas e a classificação final obtida. Desta forma, os diferentes tipos de notas foram considerados como factos, refletindo o que se pretende medir. No caso da creditação de competências apenas é considerada a nota final.

Para além da necessidade de existirem dados no ficheiro que necessitam de ser armazenados, na Tabela 6 é possível compreender a origem das dimensões e factos quando relacionados com as consultas descritas na secção de Análises. As dimensões Aluno, Ano curricular e Horário são dimensões apresentadas mas foram desenvolvidas no âmbito do *data mart* de inscrições. Inclui-se também o facto inscrições.

Tabela 6 – Matriz de relacionamento de factos e dimensões.

Factos Dimensões	Nota de frequência	Nota de exame normal	Nota de exame de recurso	Nota de exame especial	Nota Final	Inscrições
Aluno	H	I	I	I	A,B,C,D,E,F,G,J,K	A,B
Disciplina	H				G	
Regine					E	
Curso	H	I	I	I	A,B,C,D,E,F,G,J,K	A,B
Data	H	I	I	I	A,B,C,D,E,F,G,J,K	A,B
Tipo de classificação					A,B,C,D,E,F,G	A,B
Ano curricular					C,D,E,F	
Horário					F	

3.5 Arquitetura da solução

A arquitetura dimensional defendida por *Ralph Kimball* foi escolhida para o desenvolvimento da solução. O objetivo de desenvolvimento será armazenar e tratar apenas conjuntos de dados relacionados com avaliações dos alunos, o que significa que apenas será criado um *data mart* para conter este domínio de informação (estratégia *bottom-up*). Assim, os custos associados aos desenvolvimentos são significativamente reduzidos, mantendo-se sempre a possibilidade de acrescento de novas áreas de informação ao longo do tempo que possuem significativamente o mesmo custo (solução escalável). Um exemplo é o carregamento de dados relativo às inscrições dos alunos, que será desenvolvido o *data mart* específico pelo colega Tiago Pereira, aluno de MEI já enunciado.

Por outro lado, um fator importante a ter em conta é a dimensão dos dados em questão. Dado que o departamento atualmente contém um número de alunos elevado, o conjunto de informação de cada ano, por cada aluno inscrito, cresce com grande velocidade à medida que vai obtendo as classificações das provas e de novas inscrições. Desta forma, é necessário ter em conta que todo este armazenamento não prejudique o acesso à informação. Como foi descrito no capítulo anterior, esta arquitetura permite a utilização de modelos dimensionais onde normalmente o relacionamento entre dados é baixo, permitindo assim obter uma menor complexidade e rapidez no acesso, dado o número reduzido de *joins* necessários e sobretudo de fácil manutenção.

No desenvolvimento desta arquitetura vão ser utilizadas tecnologias *Microsoft*, nomeadamente o *SQL Server* para criar os modelos de armazenamento de dados, o *Microsoft Integration Services* para criar todo o *workflow* de carregamento e manutenção do AD e o *Microsoft Analysis Services* para criação do cubo de dados, com o objetivo de representar algumas

análises a partir dos dados armazenados. No que toca à apresentação dessas análises, vai ser utilizada a ferramenta *Microsoft Power BI*.

Para além das vantagens que estas ferramentas trazem para o desenvolvimento e manutenção da solução, foram escolhidas fundamentalmente pelo facto de serem disponibilizadas de forma gratuita. Dado que o Departamento de Engenharia Informática possui um acordo com a Microsoft, algumas das ferramentas são disponibilizadas para uso de forma livre por parte dos docentes e alunos.

Depois de definidas as ferramentas necessárias, na Figura 26 é apresentado o esquema geral da solução, englobando a extração, o carregamento/limpeza e acesso por parte do utilizador final.

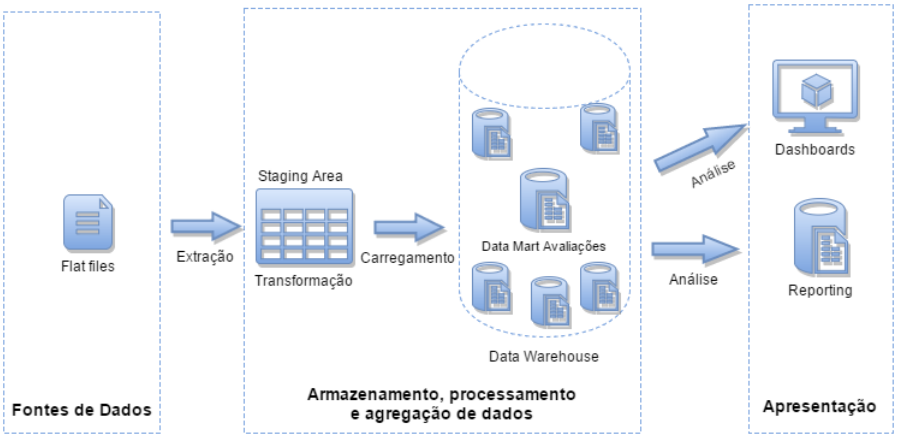


Figura 26 – Arquitetura da solução.

3.5.1 Staging Area

Numa primeira abordagem, depois se ser feita a extração, torna-se necessário armazenar os dados com o objetivo de os tratar antes de serem carregados para o sistema de armazenamento final. Assim, e tendo em conta a informação disponibilizada, foi criada a estrutura de armazenamento, apresentada na Figura 27. É de salientar que a tabela Aluno é representada de cor diferente apenas pelo facto de fazer parte do desenvolvimento do *data mart* de inscrições.

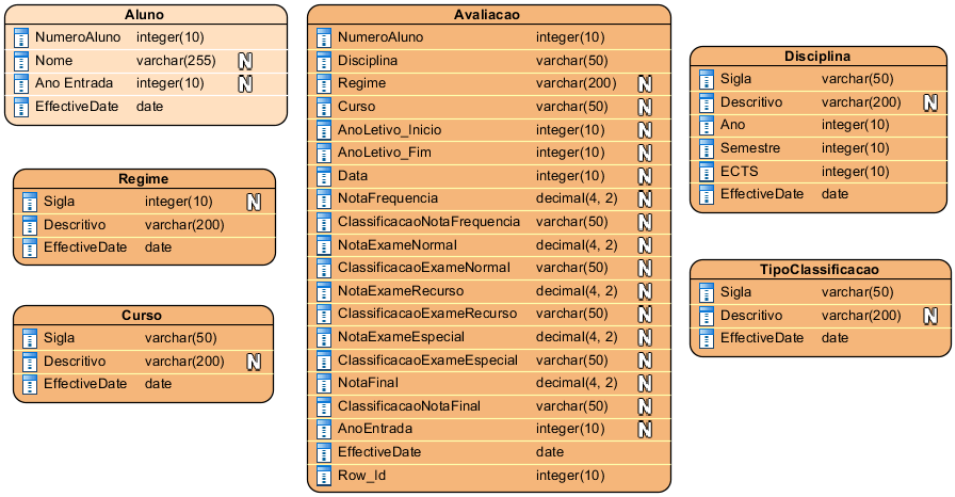


Figura 27 – Modelo de dados da área de *staging*.

Dado que a granularidade a ter em conta para este processo de negócio é referente aos registos de avaliações de alunos, foi criada a tabela Avaliação para armazenar os registos transacionais. Contém assim toda a informação até aqui mencionada, que identifica um registo de avaliação de um aluno a determinada unidade curricular. É de salientar que esta tabela contém o campo "Row_Id" com o objetivo de auxiliar no processo de deteção de duplicação de registos no decorrer do fluxo de tratamento e limpeza.

Para além desta tabela, considerou-se a existência de um conjunto de informação que se relaciona e que confere mais detalhe no domínio em específico em que está armazenada, sendo em grande parte constante, ou seja, que muito dificilmente sofre alterações. Para este caso optou-se por carregar a informação uma única vez no sistema, de forma permanente e com base nos valores possíveis existentes, que se relaciona evidentemente com os dados extraídos. Este tipo de tabelas são as seguintes:

- **Regime:** possui informação sobre os regimes que um aluno pode estar inscrito. Tem como valores possíveis: Erasmus, Extracurricular, Extraordinário, Gratuito, Integral, Internacional-Integral, Mobilidade, Parcial e Vasco da Gama;

- **TipoClassificacao:** armazena os tipos de classificações que se podem obter em unidades curriculares, excetuando a nota final numérica;

- **Curso:** entidade para armazenar os cursos disponíveis e sobre o qual o aluno está inscrito. Esta tabela foi criada com o intuito de permitir no futuro que sejam analisadas informações sobre outros cursos. Neste contexto apenas foi inserido o curso LEI;

- **Disciplina:** estrutura que armazena os dados sobre as disciplinas. Foi possível extrair a sigla, um descritivo da disciplina, o ano curricular em que se insere, o semestre e os ECTS.

Por fim restam apenas a entidade Aluno definida com o objetivo de armazenar os dados básicos dos alunos.

Associadas às tabelas apresentadas, foi também criada uma área denominada por DQP (*Data Quality Problems*) pensada para armazenar todos os registos problemáticos que surjam no decorrer do processamento. Desta forma consegue-se identificar mais facilmente problemas existentes nas fontes de dados, podendo ser validados por outrem e carregados novamente no sistema. Cada tabela existente na área de *staging* possui a sua tabela DQP associada. Este modelo também inclui uma tabela DQP para a tabela de avaliações, utilizada para armazenar a possibilidade de registos inválidos (apesar de se considerar muito reduzida) no carregamento do AD. Em conjunto com o registo inválido, caso o erro seja detetado a partir de uma validação feita ao registo no decorrer do fluxo, é registado o componente do fluxo de dados que o identificou como inválido ("component") e uma mensagem relacionada com a validação que foi feita ("DQP"). No caso em que aconteceu uma exceção no decorrer do processo, é armazenado o código do erro se existir ("ErrorCode") e o descritivo do erro ("ErrorDesc"). Nos dois casos, se for possível identificar a coluna problemática, esta também é identificada ("ErrorColumn"). De seguida apresenta-se o modelo descrito, na Figura 28.

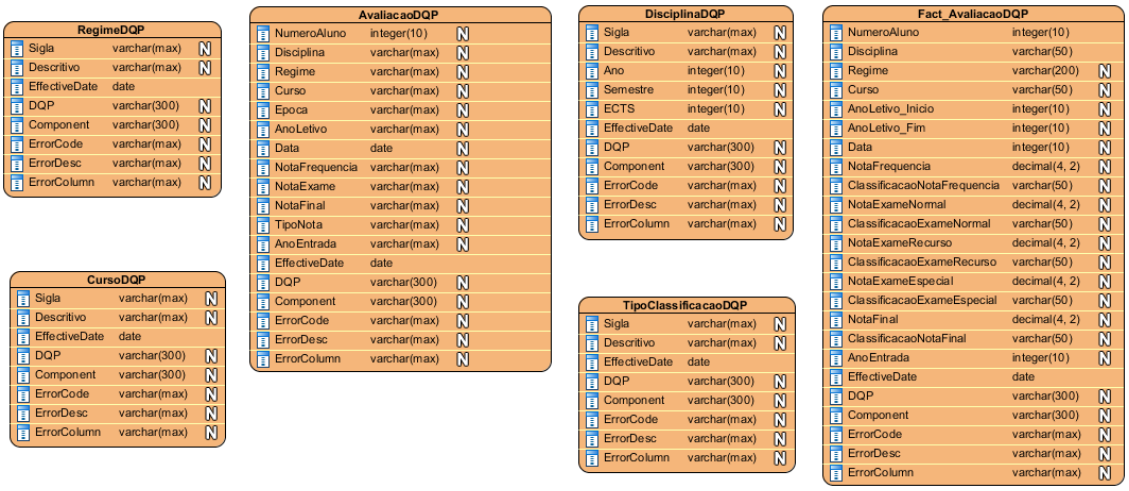


Figura 28 – Modelo de dados da área de *staging* com tabelas DQP.

O modelo desenvolvido apresenta-se sem relacionamentos entre estruturas, de forma a facilitar o carregamento dos dados, sem qualquer tipo de dependência e devido ao facto que cada uma apenas atua como tabela de *log* de registos problemáticos.

Neste modelo, existem duas tabelas de avaliação para armazenar registos inválidos. A tabela "AvaliacaoDQP" é responsável por armazenar registos inválidos no fluxo de dados de extração, antes de serem armazenados em *staging*. Por outro lado, a tabela "Fact_AvaliacaoDQP" armazena possíveis registos inválidos encontrados quando é feita a procura do registo pela respetiva chave na dimensão, no fluxo de dados de carregamento do armazém de dados. Todas as restantes tabelas armazenam registos inválidos da dimensão respetiva. Foi tomada a decisão de criar este tipo de tabelas, principalmente a "AvaliacaoDQP" dado que, numa primeira análise,

foram identificados alguns registos com números de aluno inválidos. Assim, estas tabelas permitem despistar à primeira vista o tipo de incoerências que cada registo possui.

É de salientar que, no caso deste modelo, a tabela Aluno não é representada com a respetiva DQP, dado não ser desenvolvida neste contexto, pelo facto de ter sido criada no âmbito de outro *data mart* (inscrições de alunos).

3.5.2 Armazém de dados

Depois da extração da informação, de seguida é feita uma limpeza e tratamento dos dados existentes antes de serem transportados para o AD. Foi criado o modelo dimensional (Figura 29) para armazenar esse conjunto de dados, baseado na topologia em estrela. É de salientar que a tabelas Dim_Aluno e Dim_Date são representadas de cor diferente apenas pelo facto de fazerem parte de outra dissertação. No entanto, são utilizadas neste *data mart* relacionando-se com a tabela de factos de avaliações.

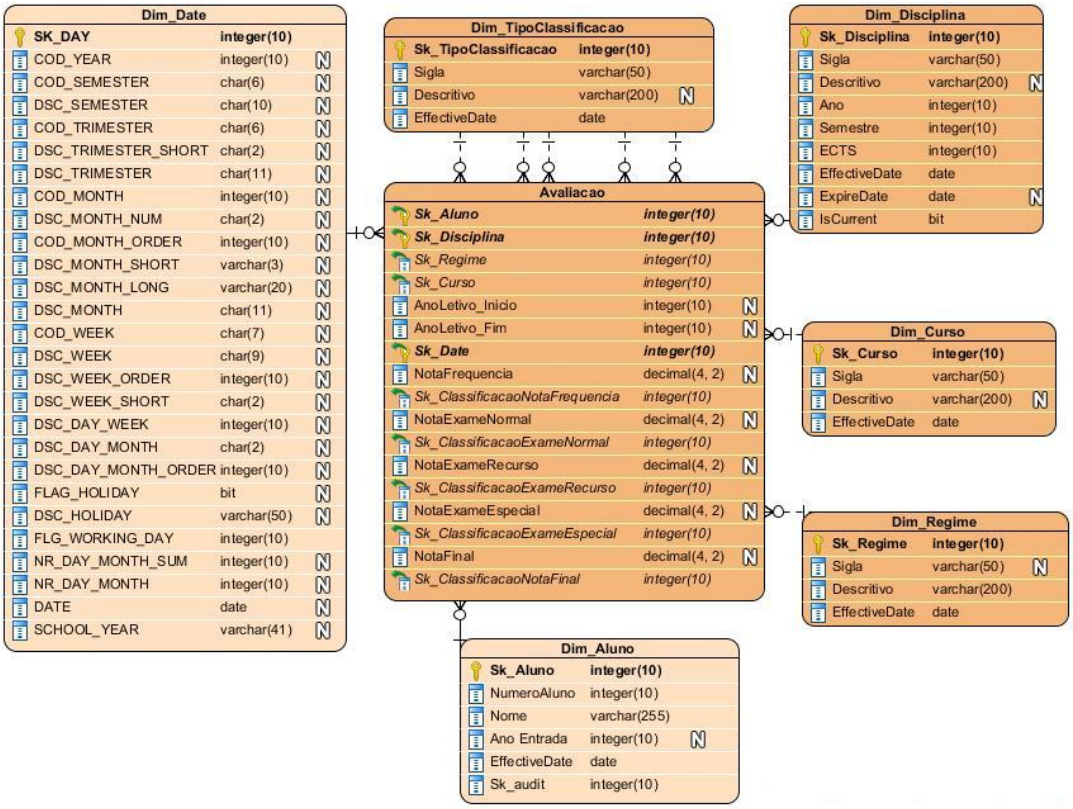


Figura 29 – Modelo dimensional do *data mart*.

Verifica-se que as estruturas das dimensões encontram-se muito semelhantes ao apresentado na camada intermédia, acrescentando-se apenas campos para manutenção de histórico, a respetiva chave numérica e a criação de um campo de data comum a todas as tabelas, com o objetivo de identificar quando é que o registo foi carregado para o sistema. Adicionalmente e

fora deste âmbito, foi criada uma dimensão data capaz de abranger as necessidades de tipos de data presentes. Optou-se por uma granularidade diária.

Relativamente à tabela de factos, a tabela Avaliação regista todas as ocorrências de avaliações existentes. Definiu-se que um registo de avaliação é identificado como único no sistema apenas a partir da conjugação do aluno, disciplina e data. Nestas condições, um aluno não pode ter dois registos iguais para a mesma disciplina, para a mesma data, sendo esta considerada como a data de obtenção da classificação ou data do ano letivo em que se insere. Assim, o aluno é associado a uma disciplina em que obteve a avaliação, a época, o curso, o regime, com um determinado tipo de avaliação e respetiva classificação final. Assim, foi encontrada aqui a necessidade de medir as avaliações por tipo de avaliação: se é uma nota de frequência, exame da época normal, exame de recurso, exame especial ou da nota final. Existem dois tipos de campos possíveis a serem preenchidos para cada tipo de avaliação, dependendo do tipo de nota: numérica ou classificação por sigla.

As Dimensões Aluno e Data não vão ser descritas ao detalhe neste contexto, pelo facto de terem sido criadas no âmbito de outro *data mart* (inscrições de alunos).

3.5.3 Estruturas auxiliaries ao processo

Com o objetivo de armazenar alguns dados relacionados com o fluxo de processamento propriamente dito, foi criada uma base de dados relacional (Figura 30) que armazena dados sobre os resultados de cada carregamento efetuado (tabela "Audit") e, com o objetivo de processar cada um dos pacotes de tratamento e carregamento sem a necessidade de executar cada um manualmente, contém informações sobre cada pacote de processamento existente (tabela "*Package*"), a que *data mart* pertence (tabela "*Star*") e a respetiva configuração de execução, permitindo associar cada pacote ao carregamento específico de um *data mart* (tabela "*PackageStar*").

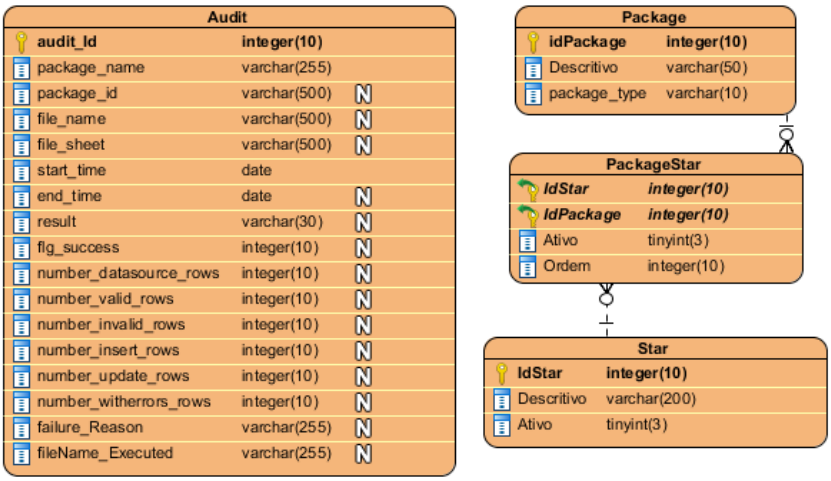


Figura 30 – Modelo da base de dados de configurações.

Quando o carregamento de um *data mart* em específico termina, são armazenados todos os dados relativos ao resultado obtido. Nesse registo é considerado o pacote que foi executado, qual o nome da folha e do ficheiro que foram utilizados para o carregamento da estrutura, a data de início e a de fim de execução, o tipo de resultado, mensagem de erro (caso exista) e a contabilização dos registos por categorias. Desta forma é possível ter sempre um registo do que aconteceu em cada execução, podendo mais tarde ser utilizada para tratamento (por exemplo: envio de informação por email para o responsável).

Por outro lado, as restantes tabelas foram desenhadas no âmbito da configuração de execução dos pacotes, permitindo que o carregamento de cada uma das tabelas de *staging* e dimensões/tabela de factos seja automatizado. Assim, inicia-se a execução de um único pacote que consulta estas tabelas (consoante a informação que foi inicialmente parametrizada nas variáveis) obtendo uma lista de pacotes pela ordem que devem ser executados.

3.6 Avaliação de Resultados

Neste contexto, pode-se definir avaliação de resultados como um conjunto de parâmetros definidos com o objetivo de serem avaliados/testados aquando do carregamento dos dados no sistema e na respetiva geração das análises. Desta forma, torna-se possível comparar casos particulares e retirar conclusões sobre qual poderá ser a melhor abordagem em concordância com o contexto. Para começar e definir então o processo de avaliação, é necessário definir as possíveis hipóteses que se pretendem alcançar, o conjunto de condições que vão ser testadas por passos na fase de experimentação, a necessidade de equipamento necessário e por fim o plano de análise de dados (Ferreira 2015).

Depois do tratamento e carregamento de dados no sistema, o grande objetivo e a hipótese neste contexto passou por verificar a validade dos mesmos e garantir que não existiram problemas a nível de cálculos ou perda de informação. A nível de medidas utilizadas, as grandezas utilizadas neste processo são a satisfação do utilizador na utilização de ferramentas de análise/criação de novos relatórios de informação e os resultados obtidos (quando comparados com as análises disponibilizadas e considerando sempre a existência de um desvio considerável).

Relativamente aos resultados obtidos do carregamento e tratamento dos dados, os testes para comprovar a veracidade do sistema passaram por comparar os valores originais de análises feitas de forma manual com a análise resultante automática. Este teste foi efetuado mais do que uma vez, considerando fontes de dados de anos letivos diferentes.

Por último, apesar da amostra de utilizadores ser restrita ao Diretor da LEI e/ou eventualmente ao Diretor do MEI, poderiam ser feitos questionários aos mesmos sobre a facilidade e a usabilidade com que utilizam os dados e criam novas análises. No entanto, dada a amostra reduzida, não possui qualquer relevância estatística. Desta forma, não é possível a realização de análises estatísticas sobre este contexto.

4 Implementação

No decorrer deste capítulo serão discriminadas as fases que englobam todo o desenvolvimento da solução. A primeira secção diz respeito não só à criação da área de *staging* que serve como base para a limpeza, transformação e armazenamento temporário dos dados, bem com a criação do AD. A seguir, serão descritas as estruturas de dados criadas para auxílio na realização de análises: cubo de dados e as tabelas agregadas. Adicionalmente, no Anexo B são especificadas as ferramentas utilizadas e a configuração do ambiente de trabalho partilhado. Será descrito com pormenor como todo este processo foi desenvolvido

4.1 Criação da base de dados de staging

Numa primeira fase foi criada toda a estrutura de *staging*, permitindo desenvolver e testar a primeira etapa do processo. Este pré-armazém de dados é caracterizado por tabelas sem qualquer tipo de relação física entre dados, isto é, não possui relações típicas de um modelo relacional com o objetivo de facilitar o processo de carregamento, quer a nível de tratamento de erros/dependência dos dados, quer a nível de performance no acesso. De seguida são apresentadas, com mais detalhe, cada uma das tabelas criadas. De notar que todas possuem um campo de data, com o objetivo de identificar em que data é que o registo foi carregado.

A partir da folha de cálculo apresentada anteriormente sobre as configurações de disciplinas (extraída das fontes de dados disponibilizadas), foi identificada a necessidade de armazenar toda a informação sobre as disciplinas numa única estrutura. Assim, a tabela representada na Figura 31 permite armazenar as iniciais da disciplina, um descritivo associado, o ano em que é lecionada, o semestre e os respetivos ECTS obtidos aquando da aprovação na unidade curricular.

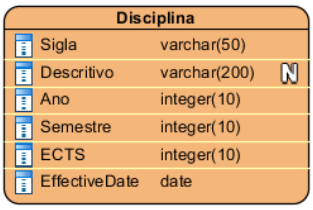
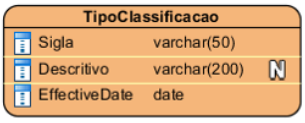


Figura 31 – Tabela Disciplina.

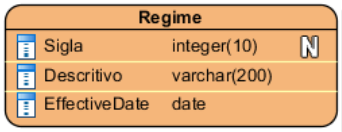
Para cada uma destas disciplinas existem diferentes tipos de nota que podem ser obtidos nos diferentes momentos de avaliação. Analisando as colunas de classificação do ficheiro fornecido (nomeadamente as colunas "NFreq", "Exame" e "Nota") consegue-se retirar a classificação obtida pelo aluno como uma nota numérica ou um tipo de classificação escrita. Com o objetivo de categorizar que tipos de classificação são válidos para cada tipo de avaliação, foi criada uma tabela (Figura 32) para armazenar a sigla que a representa e o respetivo descritivo. Para não sobrecarregar o processo de carregamento com uma validação nas colunas do ficheiro sobre os tipos de classificação diferentes possíveis, esta foi carregada com base num ficheiro *standard*, dado que este tipo de informação dificilmente se altera ao longo do tempo.



Sigla	Descritivo
NF	Não frequentou
NC	Não classificado
FT	Faltou
DT	Desistiu
SMS	Sem nota mínima Setembro
SMNF	Sem nota mínima e não frequentou
SMR	Sem nota mínima recurso
APROV	Aprovado
REPRV	Reprovado
CC	Creditação de Competências
CCP	Certificado de Competências Pedagógicas
FNS	Equivalência por Formação Não Superior

Figura 32 – Tabela tipo de classificação e respetiva folha de carregamento.

O mesmo se aplicou à tabela de regimes (Figura 33). Foi criada uma folha adicional neste ficheiro de dados estático com o objetivo de carregar todos os regimes existentes, considerando aqui uma sigla definida para futuramente facilitar nas possíveis análises a criar.

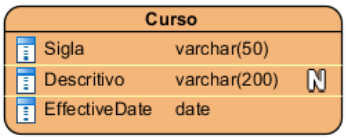


Sigla	Descritivo
E	Erasmus
ExtrCurr	Extra-Curricular
Extr	Extraordinário
G	Gratuito
I	Integral
II	Internacional-Integral
M	Mobilidade
P	Parcial
VC	Vasco da Gama

Figura 33 – Tabela Regime e respetiva folha de importação de dados.

Como anteriormente referido, a informação curricular refere-se apenas aos alunos que frequentaram a LEI. Assim, para armazenar a informação sobre o curso, foi criada a tabela

representada na Figura 34 para armazenar o nome do curso e a respetiva sigla, permitindo que futuramente seja possível alargar a outros curso da instituição.



Sigla	Descritivo
LEI	Licenciatura em Engenharia Informática

Figura 34 – Tabela Curso e respetiva folha de importação de dados.

Por último, resta a estrutura de dados para armazenar os dados transacionais (Figura 35). Armazena registos de avaliação de um determinado aluno, a disciplina que frequentou, regime de inscrição, num determinado ano e, caso exista informação, em que data foi obtida a classificação. Associado, tem-se a avaliação propriamente dita, que pode ser composta por classificações numérias ou classificações textuais e de diferentes épocas/fases: nota de frequência obtida no decorrer do ano letivo, nota de exame da época normal, nota de exame da época de recurso, nota de exame da época especial e a nota final obtida. Quando a classificação obtida corresponde a uma sigla, utiliza-se a chave da respetiva dimensão. Caso contrário, a nota numérica obtida é armazenada como sendo uma medida.

Nesta estrutura foi adicionado um índice de linha, com o objetivo de detetar duplicados no decorrer do fluxo de dados. Todo o processo associado a este campo vai ser descrito no passo seguinte.

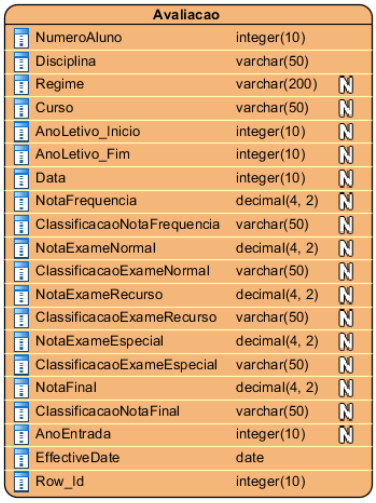


Figura 35 – Tabela de Avaliação.

4.2 Processo ETL

Nesta secção é descrito de que forma foi implementado todo o processo de extração, limpeza e carregamento dos dados. Este, começa a partir de um pacote criado que executa os pacotes de processamento de forma automática. Numa primeira fase, é feita uma procura nas tabelas

de configurações pela lista de pacotes a executar e respetiva ordem de execução. É possível parametrizar o tipo de carregamento, definindo o valor de uma variável local com o valor respeitante ao tipo de execução que se pretende:

- Por *data mart*, pode ser preenchida com o valor "Avaliação" para executar todos os *packages* necessários e que digam respeito apenas ao carregamento de informação sobre as avaliações dos alunos;

- Por área, executando a área de *staging* ("STG"), dimensões ("DIM"), as tabelas de factos ("FCT") ou atualização do cubo ("CUBE");

- Execução relativa a todos os *data marts* existentes, utilizando o valor "*Complete*" para executar todos os *packages* do projeto;

- Execução apenas relativas a tabelas que armazenem informações sobre cada evento do processo de negócio (neste caso pacotes de execução que envolvam as tabelas de factos e as tabelas de *staging* associadas), utilizando o valor "Parcial".

Na Figura 36 é possível verificar a lista de execução devolvida para cada tipo de parametrização.

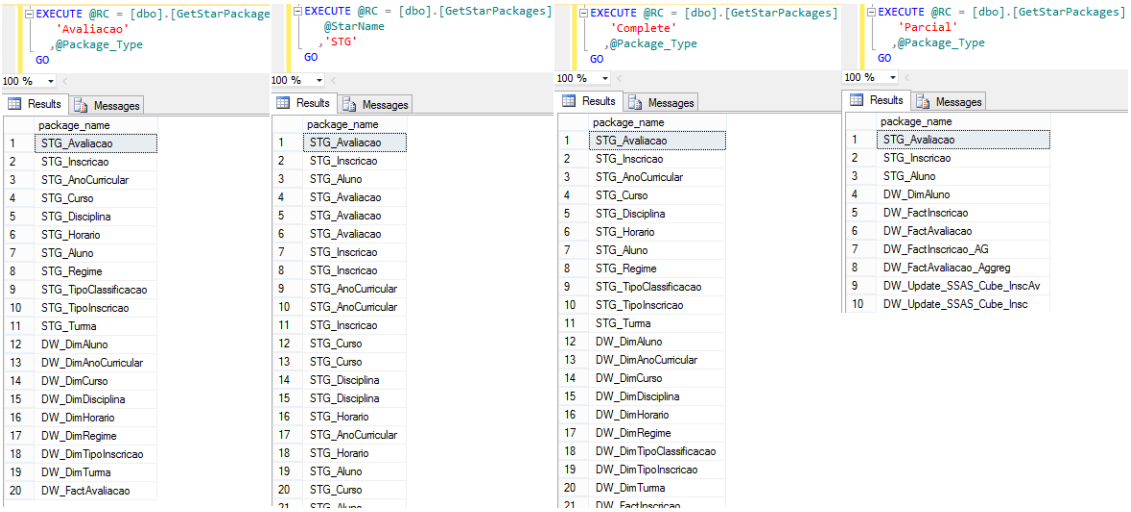


Figura 36 – Tabelas de configuração e tipos de parametrização possíveis.

Depois da obtenção da lista de itens a executar, é feito um ciclo para executar cada pacote de processamento. Quando o processamento de todos os pacotes termina, é captada a data/hora de fim de execução e é enviado um email com o resultado desde a data de início de execução até à data final capturada, com base nos registos existentes na tabela de auditoria. É construída uma tabela em HTML, que é enviada através do comando *SQL server* "sp_send_dbmail" (Figura 37).

```
OPEN cursor_failure_reason;
FETCH NEXT FROM cursor_failure_reason INTO @TimeStamp, @package_name, @number_datasource_rows, @number_witherrors_rows, @failure_Reason
While @@FETCH_STATUS = 0
BEGIN
     SET @count = @count +1

     SET @content_table_content = ISNULL(@content_table_content, '') +
                                 '<tr>
                                       <td align="center">' + CAST(@TimeStamp AS varchar(20)) + '</td>
                                       <td align="center">' + @package_name + '</td>
                                       <td align="center">' + CAST(@number_datasource_rows AS NVARCHAR(300)) + '</td>
                                       <td align="center">' + CAST(@number_witherrors_rows AS NVARCHAR(300)) + '</td>
                                       <td align="center">' + ISNULL(@failure_Reason,'') + ' (...)</td>
                                  </tr>'

        --SET @content_table_content +=  ISNULL(@content_table_content, '') + ' 1 '

    FETCH NEXT FROM cursor_failure_reason INTO @TimeStamp, @package_name, @number_datasource_rows, @number_witherrors_rows, @failure_Reason
END

CLOSE cursor_failure_reason;
DEALLOCATE cursor_failure_reason;


SET @STR_CONTENT = '<HTML>
<head></head>
<body>
   <p>Some errors occurred shile the etl process is running.</p>
   <p>See the details below.</p>
   <table border="2">'
     +@content_table_header+
   ''+@content_table_content+
'</table></body></HTML>';

EXECUTE msdb.dbo.sp_send_dbmail @profile_name = 'DEV_TEAM', @recipients = 'giselacouto@live.com.pt',
 @subject = 'ETL Process execution', @body = @STR_CONTENT, @body_format = 'HTML';
```

Figura 37 – Extrato de código de envio de correio eletrónico.

De notar que, esta parte foi desenvolvida em conjunto com o colega Tiago que desenvolveu um *data mart* relativo às inscrições dos alunos. Assim, o processo de execução no geral fica uniformizado para qualquer tipo de *data mart* que existe e que possa vir a existir no futuro.

4.2.1 Staging Area

De uma forma geral, todos os processos de extração de dados criados são semelhantes. Na área de controlo do fluxo do processo (*Control Flow* do MSSDT) primeiramente foi feita uma limpeza à tabela de *staging*, incluindo também a tabela correspondente de armazenamento de registos com problemas de qualidade de dados. O componente utilizado é um componente do tipo "*Execute SQL TASK*", contendo apenas um *script* de limpeza, com o objetivo de começar o processo de carregamento com as tabelas sem dados.

O próximo passo deste fluxo é composto por um componente do tipo *Data Flow Task*, onde a transformação e o carregamento de dados acontece, começando pela definição da fonte dos dados e das colunas a extrair. Para cada tipo de ficheiro existente (fontes de dados reais e fontes de dados estáticas) e dado que a utilização dos componentes do tipo *Excel Data Source* comportavam alguns problemas de estabilidade de conexão para a solução, foram criados *linked servers* distintos com o objetivo de criar alguma estabilidade na conexão entre os ficheiros e o projeto.

Inicialmente, é feita uma *query* ao *linked server* pretendido para capturar os dados, especificando a folha e as colunas necessárias. Na Figura 38 são apresentadas as *queries* para

acesso a uma folha do ficheiro de dados fixos, do ficheiro de notas de alunos e do ficheiro de avaliações do aluno (todas as folhas/ficheiros utilizados neste âmbito).

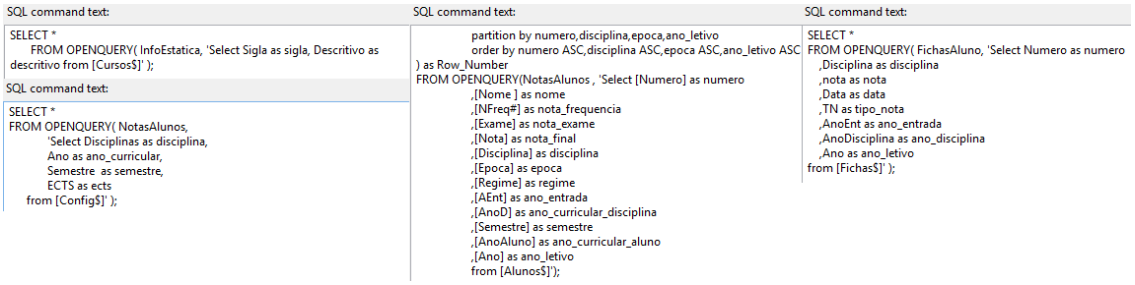


Figura 38 – Ligação da solução com os ficheiros de dados.

De seguida, foi feito um mapeamento de cada coluna existente no ficheiro para variáveis internas. Esta camada abstrata foi criada com o objetivo de permitir alterações nos componentes anteriores (nome das variáveis, tipos de dados) sem que o processo seguinte sofra grandes impactos. Este mapeamento é composto por dois passos: primeiramente os valores das colunas são extraídos para variáveis do tipo texto e de seguida tenta-se converter esses valores obtidos no tamanho e nos tipos de dados pretendidos. Adicionalmente, é sempre capturada a data atual, que vai acompanhar o registo até ser inserido ("*Effective Date*").

Caso ocorra algum erro ao tentar executar estes tipos de operações, será registada a linha onde se deu a ocorrência com a respetiva mensagem de erro fixa *"An error occurred while data row is converted"*, identificando também em que componente e em que pacote de execução aconteceu. Este erro será inserido na respetiva tabela de DQP. Por fim, os dados são inseridos na tabela de *staging*, mapeando cada uma das variáveis do *workflow* com cada coluna da tabela (Figura 39). Os erros de inserção são igualmente tratados com a mensagem *"Insert row in OLE DB Destination caused an error*".

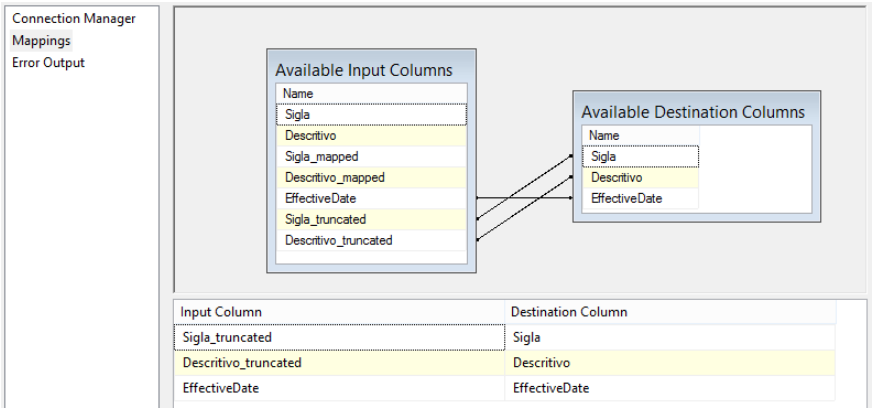


Figura 39 – Exemplo de mapeamento de variáveis para a tabela de *staging area*.

Para monitorizar a quantidade de registos que são processados por cada tabela à medida que se avança no fluxo, foram utilizados componentes de contagem para contar a totalidade de

registos da fonte, os que são válidos/inválidos, os que contêm erros específicos e o número de registos que foram inseridos/atualizados.

Quando é terminada a execução do componente de fluxo de dados, é sempre registado na tabela auditoria, o detalhe do resultado de execução do pacote e/ou carregamento da tabela. Com base em variáveis, é registado o pacote que foi executado, associado do respetivo identificador, a data em que começou e terminou a execução, o descritivo do resultado do processamento ("*Success", "Failed*"), a contagem de registos enunciada anteriormente, a falha e o nome do ficheiro que foi carregado. Desta forma, o resultado de todos os carregamentos pode ser armazenado de forma genérica para cada um dos processos como representado na Figura 40. Os componentes utilizam um procedimento para inserção dos registos na tabela, considerando todos os parâmetros anteriormente anunciados.

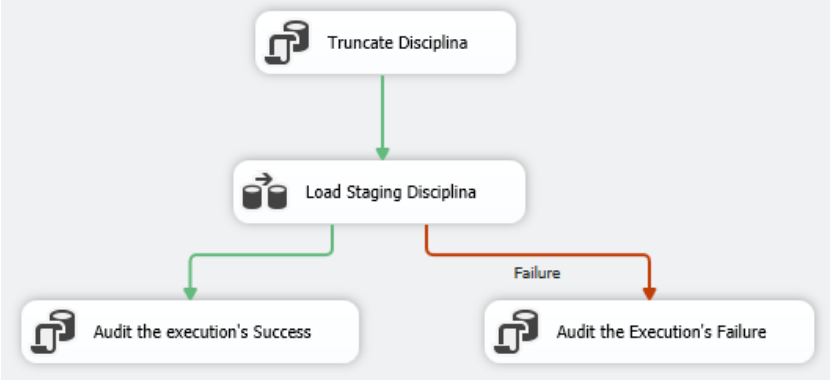


Figura 40 – Exemplo de auditoria de resultados.

Passando para uma vertente mais detalhada, de seguida são apresentadas as especificidades do carregamento de cada tabela.

Disciplina

Esta tabela tem como origem uma das folhas de cálculo fornecidas que contém apenas informações sobre todas as disciplinas (ver Fontes de Informação, Figura 24). Foi extraída a informação das seguintes colunas, sendo que numa primeira abordagem, todas as colunas foram mapeadas para o processo na sua forma mais simples (Figura 41):

- **Disciplinas:** texto com a designação de cada disciplina existente;
- **Ano:** valor numérico entre um e três;
- **Semestre:** valor numérico entre um e seis;
- **ECTS:** valor numérico de créditos.

Derived Column Name	Derived Column	Expression
Sigla_mapped	<add as new column>	disciplina
Descritivo_mapped	<add as new column>	disciplina
Ano_mapped	<add as new column>	ano_curricular
Semestre_mapped	<add as new column>	semestre
ECTS_mapped	<add as new column>	ects
EffectiveDate	<add as new column>	GETDATE()

Figura 41 – Mapeamento entre ficheiro e *workflow* de extração das disciplinas.

De seguida procedeu-se à sua conversão/transformação, convertendo os valores necessários em numéricos e truncando os campos de texto para que os limites máximos definidos não sejam ultrapassados durante o processamento, como apresenta a Figura 42.

Input Column	Output Alias	Data Type	Length
Descritivo_mapped	Descritivo_truncated	Unicode string [DT_WSTR]	200
Ano_mapped	Ano_truncated	four-byte unsigned integer ...	
Semestre_mapped	Semestre_truncated	four-byte signed integer [D...	
ECTS_mapped	ECTS_truncated	four-byte signed integer [D...	
Sigla_mapped	Sigla_truncated	Unicode string [DT_WSTR]	50

Figura 42 – Transformação de dados das disciplinas.

A partir do momento em que a sintaxe dos dados se encontra correta, procede-se à validação semântica, identificando para este caso em específico se o ano letivo, o semestre e os ECTS introduzidos são válidos (Figura 43). Foi considerado que o ano curricular teria que compreender entre um e três, o semestre entre um e sete e por fim, os ECTS entre um e sessenta.

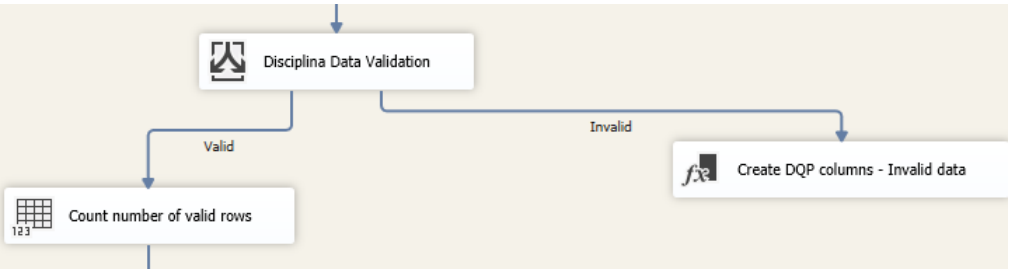


Figura 43 – Validação de dados das disciplinas extraídas.

Os dados válidos são enviados para armazenamento na tabela Disciplina da *staging area* (Figura 44). Já os registos inválidos são armazenados na tabela de DQP associada com a mensagem de erro *"Datasource row is invalid"*.

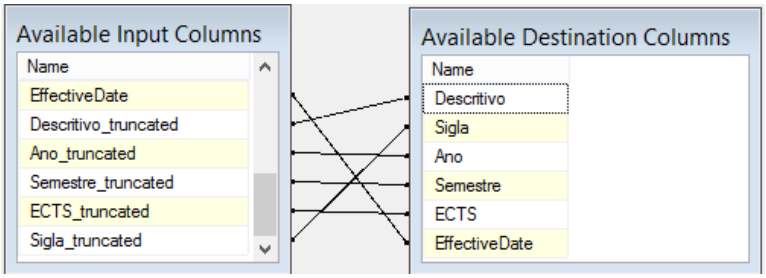


Figura 44 – Mapeamento para armazenamento na tabela de *staging* Disciplina.

Tipo de Classificação

Para carregamento desta tabela foi criada uma folha de cálculo *standard*, dado que este tipo de informação é geralmente utilizado por sistemas de ensino e não sofre mudanças (ver Criação da base de dados *de* staging, Figura 32). Os tipos de classificação são identificados nos campos de notas pelas siglas. Dado que a sigla e o descritivo associado são do tipo texto, os valores apenas foram truncados, como apresenta a Figura 45.

Input Column	Output Alias	Data Type	Length
Sigla_mapped	Sigla_truncated	Unicode string [DT_WSTR]	50
Descritivo_mapped	Descritivo_truncated	Unicode string [DT_WSTR]	200

Figura 45 – Transformação de dados dos tipos de classificação.

De seguida, e como não é possível fazer nenhuma validação em específico, os dados são armazenados na tabela de *staging* Tipo de Classificação (Figura 46). Caso ocorra algum erro, este é registado na tabela DQP associada.

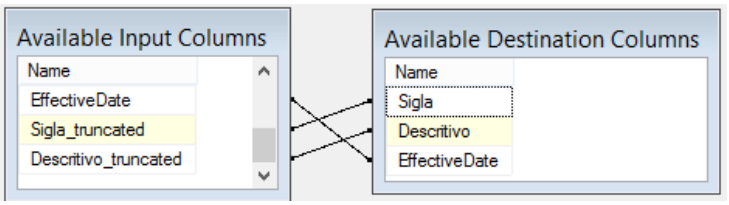


Figura 46 - Mapeamento para armazenamento na tabela de *staging* de tipos de classificações.

Regime

Tal como a tabela anterior, a fonte de dados relativamente aos regimes existentes foi também baseada em dados fixos (ver Criação da base de dados de staging, Figura 33). Nas avaliações registadas, o regime que o aluno frequentou encontra-se descriminado pelo descritivo apresentado. A nível de transformações, apenas foi limitada a quantidade de caracteres possíveis de carregar, como apresenta a Figura 47.

Input Column	Output Alias	Data Type	Length
Descritivo_mapped	Descritivo_truncated	Unicode string [DT_WSTR]	200
Sigla_mapped	Sigla_truncated	Unicode string [DT_WSTR]	50

Figura 47 – Transformação de dados de tipos de regime.

Por último, e dado que não foram feitas validações aos dados, é feito o armazenamento dos elementos válidos na tabela de *staging* Regime (Figura 48). Se por algum motivo existir problemas com registos, estes são igualmente armazenados na tabela DQP associada.

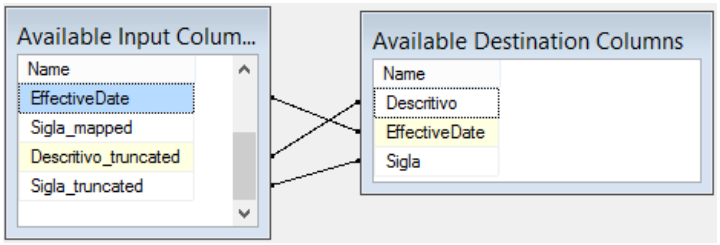


Figura 48 – Mapeamento para armazenamento na tabela de *staging* Regime.

Curso

Este tipo de informação não existe registado de forma explícita nas fontes fornecidas. No entanto, e de forma a permitir que no futuro sejam integrados novos cursos no sistema, foi definida a importação da sigla "LEI" e do descritivo do curso "Licenciatura em Engenharia Informática" (ver Criação da base de dados de staging, Figura 34). Esta informação é truncada por segurança e carregada na tabela de staging de cursos (Figura 49). Caso ocorra algum problema com algum registo, este é armazenado na DQP associada.

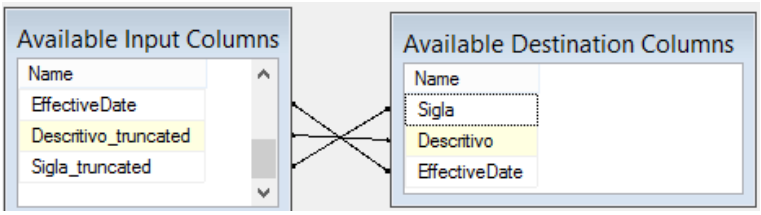


Figura 49 – Mapeamento para armazenamento na tabela de *staging* Curso.

Avaliações

Relativamente à extração e carregamento dos registos de avaliação de cada aluno, estes passam pela análise de duas fontes de informação distintas: conjunto de dados de avaliações e de classificações finais (ver Fontes de Informação, Figura 23 e Figura 25). Primeiramente, extraiu-se informação da fonte que contém o detalhe de cada avaliação obtida.

O primeiro passo no fluxo de carregamento e limpeza, passou por retirar os dados da fonte com base no *linked server* criado denominado por "NotasAlunos". É feita uma *query* que contém as colunas com informação relativa a cada avaliação e uma coluna específica que identifica de

forma unívoca, a existência de possíveis duplicados no fluxo de dados (Tabela 7). Os registos são agrupados por número de aluno, disciplina, época e ano letivo. Em cada grupo, cada registo é numerado com o seu índice respetivo, começando em um e terminando num valor que corresponde ao número de registos duplicados encontrados. Optou-se por inserir em *staging* o primeiro registo encontrado (o que tem menor índice respetivo), sendo todos os outros enviados para a tabela de DQP associada. No capítulo de Análise de resultados será possível verificar que se encontrou a necessidade deste tipo de tratamento dentro do próprio fluxo de dados, devido ao facto de se ter encontrado a existência de registos duplicados.

Tabela 7 – Exemplo de deteção de duplicados.

Índice	Número	NFreq	Exame	Nota	Disciplina	Época	Ano
1	1030093			15	PESTI	NM	20132014
2	1030093		FT	NC	PESTI	NM	20132014
3	1030093		FT	NC	PESTI	NM	20132014
1	1000051	9,5	7,7	9	ALGAV	NM	20132014
1	1000055	0		SMNF	SGRAI	NM	20132014

De seguida, na Tabela 8, apresentam-se alguns exemplos de registos de avaliação extraídos do ficheiro, considerando apenas as colunas que vão ser extraídas.

Tabela 8 – Extração exemplo da folha de cálculo Notas Alunos.

Número	NFreq	Exame	Nota	Disciplina	Época	Regime	AEnt
1000122	9,5	7,7	9	ALGAV	NM	Parcial	00
1000122		6,8	SMS	ALGAV	RE	Parcial	00
1000122	10		10	COMPA	NM	Parcial	00
1000122		FT	NC	PESTI	NM	Parcial	00
1010119	17,8	15,8	17	IARTI	NM	Integral	01
1010115	14	FT	NC	BDDAD	NM	Parcial	01
1010115	8		NC	ESINF	NM	Parcial	01

De igual forma que os processos anteriormente descritos, cada uma das colunas foi mapeada para o processo. Depois do mapeamento, foi feito um primeiro tratamento de informação (Figura 50). Salienta-se a remoção de possíveis espaços nos dados. Foram definidos alguns parâmetros fixos, dado que ainda não existem esses tipos de informação explícitos nas fontes, como o curso e os anos letivos definidos a partir de variáveis (que devem ser atualizadas aquando do carregamento para o valor pretendido). O campo Data assume o valor "-2", dado que neste ficheiro em específico não existe este tipo de informação, sendo apenas preenchido no processo de extração seguinte por um valor por defeito. Por último, a informação sobre o ano letivo corrente foi extraída, sendo repartida por dois campos distintos, com o objetivo de melhor identificar o ano de início e o de fim da atividade letiva.

Derived Column Name	Derived Column	Expression	Data Type
NumeroAluno_mapped	<add as new column>	numero	double-precision float [D...
Disciplina_mapped	<add as new column>	TRIM(disciplina)	Unicode string [DT_WSTR]
Epoca_mapped	<add as new column>	TRIM(epoca)	Unicode string [DT_WSTR]
Regime_mapped	<add as new column>	LEN(TRIM(regime)) != 0 ? TRIM(regime) : "Not Applicabl...	Unicode string [DT_WSTR]
Curso_mapped	<add as new column>	@[$Project::Curso]	Unicode string [DT_WSTR]
NotaFrequencia_mapped	<add as new column>	TRIM(nota_frequencia)	Unicode string [DT_WSTR]
NotaExame_mapped	<add as new column>	TRIM(nota_exame)	Unicode string [DT_WSTR]
Data_mapped	<add as new column>	-2	four-byte signed integer [...
AnoEntrada_mapped	<add as new column>	ano_entrada	Unicode string [DT_WSTR]
NotaFinal_mapped	<add as new column>	TRIM(nota_final)	Unicode string [DT_WSTR]
AnoLetivo_Inicio_mapped	<add as new column>	LEN((DT_WSTR,8)ano_letivo) != 0 ? SUBSTRING((DT_WS...	Unicode string [DT_WSTR]
AnoLetivo_Fim_mapped	<add as new column>	LEN((DT_WSTR,8)ano_letivo) != 0 ? SUBSTRING((DT_WS...	Unicode string [DT_WSTR]
AnoLetivo_mapped	<add as new column>	(DT_WSTR,8)ano_letivo	Unicode string [DT_WSTR]

Figura 50 – Transformação da informação da folha de notas de alunos.

O passo seguinte passou por validar o índice de cada registo para remover os registos duplicados do fluxo. Só continuam no processo todos os registos que nesse campo contenham o valor um. Caso contrário, é detetado como duplicado e é armazenado na tabela de DQP com a informação associada (Figura 51).

Figura 51 – Deteção de duplicados no fluxo de informação.

De seguida, foi feita uma verificação nas classificações obtidas (como representado no Código 2), verificando-se para cada registo de nota se é uma nota do tipo numérica ou uma classificação em forma de texto, guardando cada um dos resultados em duas variáveis distintas. Esta validação é feita ao nível das notas de frequência, notas de exame e notas finais.

```
string strNotaFrequencia =
    Row.NotaFrequenciamapped != null ? Row.NotaFrequenciamapped.Trim()
    : Row.NotaFrequenciamapped;

#region Nota de Frequencia

if (!string.IsNullOrEmpty(strNotaFrequencia))
{
    decimal resultNotaFrequencia;
```

```
    //Verificar se é uma nota numerica
    if (ConvertStringToDecimal(strNotaFrequencia, decimalSeparator, out
resultNotaFrequencia))
    {
        //Significa que estamos perante uma classificacao numerica
        Row.NotaFrequenciatrans = resultNotaFrequencia.ToString();
        Row.ClassificacaoNotaFrequenciatrans = "-2";
    }
    else
    {
        //Significa que estamos perante uma sigla
        Row.NotaFrequenciatrans = null;
        Row.ClassificacaoNotaFrequenciatrans = strNotaFrequencia;
    }
}
else
{
    Row.NotaFrequenciatrans = null;
    Row.ClassificacaoNotaFrequenciatrans = "-2";
}

#endregion
```

Código 1 – Exemplo de validação das notas de frequência.

Neste passo, também foi transformado o ano de entrada de dois dígitos para quatro dígitos. O extrato de código seguinte apresenta como foi feita essa transformação (Código 3).

```
int anoEntrada;
string anoEntradaRes = null;
string currentAnoEntrada = Row.AnoEntradamapped;

if (int.TryParse(currentAnoEntrada, out anoEntrada))
{

    if (anoEntrada >= 0 && anoEntrada < 80)
    {
        anoEntradaRes = "20" + currentAnoEntrada;
    }
    else if (anoEntrada >= 80 && anoEntrada <= 99)
    {
        anoEntradaRes = "19" + currentAnoEntrada;
    }
    else
    {
        anoEntradaRes = currentAnoEntrada;
    }
}
else
{
    anoEntradaRes = null;
}

Row.AnoEntradatrans = anoEntradaRes;
```

Código 2 – Transformação do ano de entrada do aluno.

Depois do passo de conversão de dados, foram feitas validações semânticas (Figura 52). Verifica-se se o ano letivo carregado é válido, sendo que cada ano carregado (ano letivo de início e ano letivo de fim) tem que estar ambos compreendidos entre 1900 e o ano atual e ao qual, a diferença entre os dois anos tem de ser igual a um único ano (componente "*Validate Ano Letivo*"). Ao nível do número do aluno, disciplina e classificações, se existirem valores vazios ou nulos foi considerado que não são válidos para continuarem o processo (componente "*Check record invalid fields*"). No caso da época (componente "*Check if the column Epoca is valid*"), apenas são consideradas válidas quando contém uma das seguintes siglas: "NM" (época normal), "RE" (época de recurso) e "ES" (época especial). Este campo vai ser utilizado mais à frente para se conseguir distinguir cada tipo de nota (por exemplo se é nota de exame normal ou nota de exame de recurso).

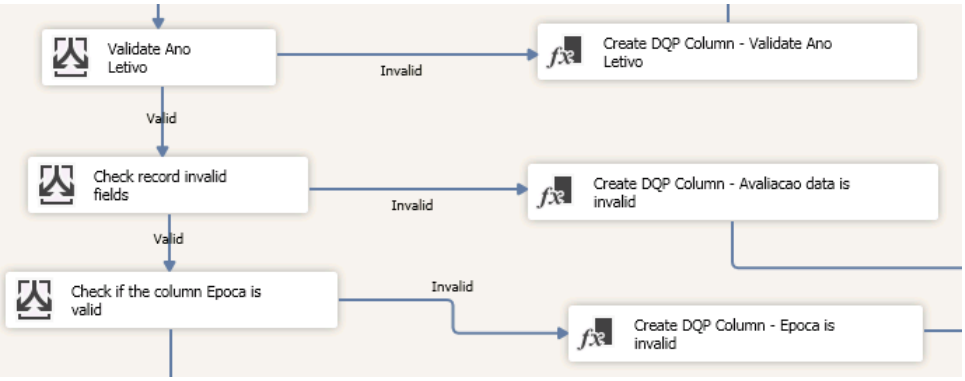


Figura 52 – Validações resultantes da extração e conversão de dados de avaliações.

Encontrados os dados válidos, estes são triados pela época de obtenção da classificação com a finalidade de construir um registo único que englobe todos os registos de classificação para o aluno e para a disciplina (naquele ano letivo, com a nota de época de exame normal, recurso e especial). Primeiramente, foi criado um fluxo de dados para cada época, colocando os valores de classificações deduzidos no processo anterior (nota de frequência, exame e final) numa variável nova de acordo com a época em específico: classificação do exame normal, classificação do exame de recurso, classificação do exame especial (subdividido por nota numérica e sigla de tipo de classificação). A Figura 53 apresenta este mesmo fluxo.

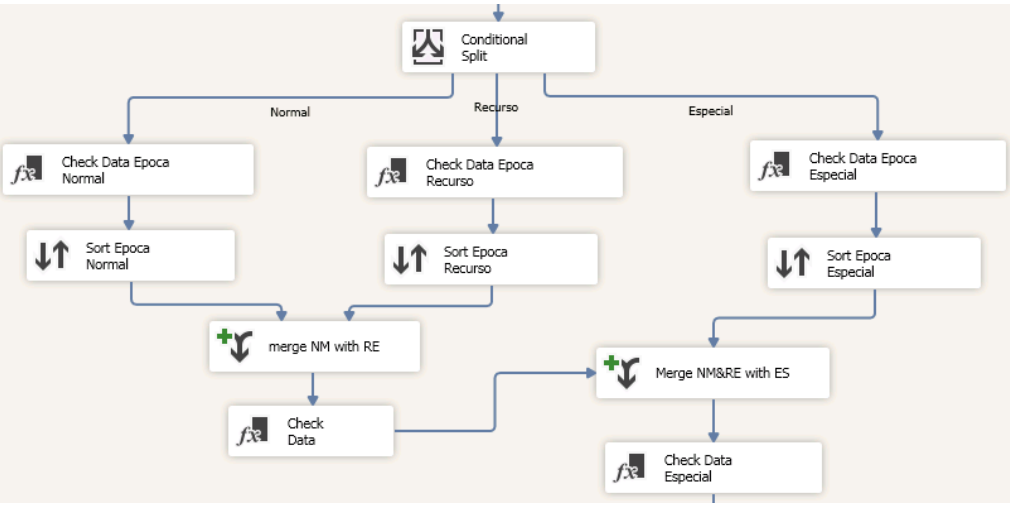


Figura 53 – Divisão do fluxo de dados por época.

De seguida, os registos são ordenados de forma ascendente (componentes "*Sort*") por número de aluno, disciplina e ano letivo. Desta forma, os dados são agrupados de forma mais facilitada e em diferentes fases, começando por juntar a informação relativa à época normal e à época de recurso ("merge NM with RE"). Optou-se por fazer um *left outer join* entre os dois fluxos de dados, procurando registos iguais no fluxo de dados de recurso pelo número de aluno, pela disciplina e pelo ano letivo, construindo um único registo a partir das notas de época normal e recurso obtidas. Na Figura 54 é apresentado como foi construído o componente para permitir a junção de registos apenas num.

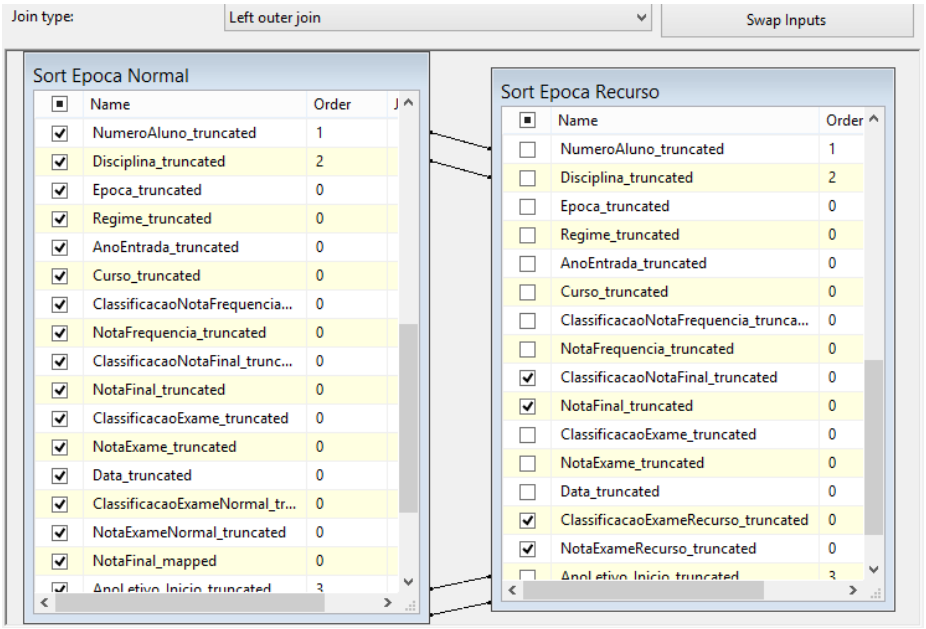


Figura 54 – Agregação dos dados de classificações de época normal e de recurso.

Para além das notas resultantes de exames e frequências nas aulas, as fontes de dados também possuem classificações finais obtidas no final de todas as avaliações possíveis daquele

semestre/ano letivo. Neste passo, a última nota obtida foi guardada não só como nota obtida na classificação (normal ou recurso) mas também como nota final, dado que neste ponto ainda não se sabe se existirão outras notas sobre o registo em questão (ver exemplo Tabela 8, onde é possível verificar que o ficheiro contém mais do que um registo de avaliação por disciplina e por aluno no mesmo ano letivo). Para este caso, tem-se como exemplo quando o aluno realiza um exame normal e de seguida um exame de recurso, sendo que para os casos em que não conseguiu obter aprovação existe outro registo na fonte de dados para a avaliação que o aluno tentou repetir, cujo campo época é diferente. De notar que se a nota do exame que tem classificação for preenchida por uma classificação textual, a nota final também assumirá esse valor textual.

Desta forma, e depois de concluída a junção de dados, são feitas novas validações. É determinada a nota final neste conjunto, dependendo do preenchimento das notas finais auxiliares de cada época, quer a nível numérico, quer a nível de tipo de classificação (é considerada como nota final a classificação que tiver valor). No caso das classificações textuais, é considerado o valor "-2" quando são vazias (para facilitar no mapeamento posterior, explicado mais à frente no carregamento do AD).

De seguida, foi feita uma nova junção entre os dados resultantes e as notas relativas à época especial (Figura 55), permitindo adicionar no mesmo registo informação sobre esta época de exame. Foi utilizado o mesmo tipo de junção descrito anteriormente, alterando-se apenas o preenchimento de uma nova coluna no registo. Após este processo são feitos os mesmos tipos de validações, alterando-se a nota final caso exista alguma classificação nesta época.

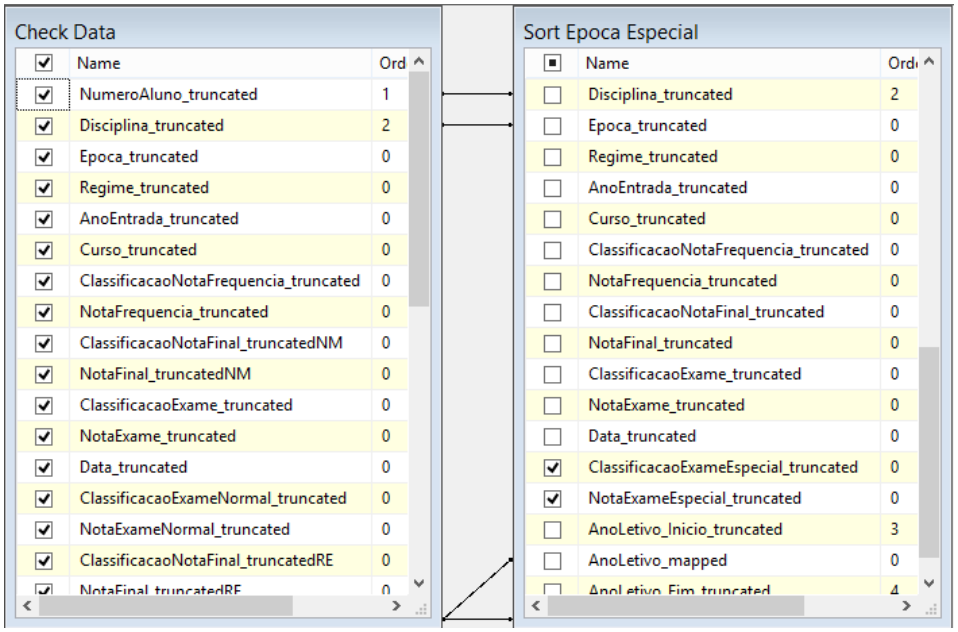


Figura 55 - Agregação dos dados de classificações de época normal, recurso e especial.

Antes dos registos serem armazenados na tabela de *staging*, verifica-se se já existe classificação final textual para o registo em questão. Se não existir (ou seja, a nota final é numérica) e dado

que este tipo de informação não existe no ficheiro facultado, calcula-se se o aluno obteve ou não aprovação à disciplina. Nesta fase assumiu-se que o aluno seria aprovado a partir do momento em que a sua nota final seja igual ou superior a nove vírgula cinco valores e, o caso contrário reflete que o aluno reprovou. Voltando ao fluxo de dados, como para cada nota numérica existe um campo de classificação associado, neste contexto este campo pode tomar os valores "APROV" e "REPROV". Por fim, o registo é inserido na tabela de *staging* de Avaliações, com o mapeamento apresentado na Figura 56.

Input Column	Destination Column
NumeroAluno_truncated	NumeroAluno
Data_truncated	Data
NotaFrequencia_truncated	NotaFrequencia
NotaExameNormal_truncated	NotaExameNormal
NotaExameRecurso_truncated	NotaExameRecurso
NotaExameEspecial_truncated	NotaExameEspecial
NotaFinal_truncated	NotaFinal
AnoLetivo_Inicio_truncated	AnoLetivo_Inicio
AnoLetivo_Fim_truncated	AnoLetivo_Fim
AnoEntrada_truncated	AnoEntrada
EffectiveDate	EffectiveDate
Disciplina_truncated	Disciplina
Regime_truncated	Regime
Curso_truncated	Curso
ClassificacaoNotaFrequencia_truncated	ClassificacaoNotaFrequencia
ClassificacaoExameNormal_truncated	ClassificacaoExameNormal
ClassificacaoExameRecurso_truncated	ClassificacaoExameRecurso
ClassificacaoExameEspecial_truncated	ClassificacaoExameEspecial
ClassificacaoNotaFinal_truncated	ClassificacaoNotaFinal

Figura 56 - Mapeamento para armazenamento na tabela de *staging* Avaliação.

O processo seguinte passa por tratar a informação do ficheiro de classificações finais. Nesta fonte de dados apenas se tem a informação do aluno, disciplina, o ano de entrada, o ano letivo correspondente, a classificação final obtida, respetiva data de obtenção e o tipo de nota, identificando se a nota foi obtida por creditação de competências. A informação destas colunas é convertida para o formato correto, como apresenta a Figura 57.

Derived Column Name	Derived Column	Expression
NumeroAluno_mapped	<add as new column>	numero
Disciplina_mapped	<add as new column>	TRIM(disciplina)
Data_mapped	<add as new column>	data
NotaFinal_mapped	<add as new column>	nota
TipoNota_mapped	<add as new column>	TRIM(tipo_nota)
Ano_Data_mapped	<add as new column>	YEAR(data)
AnoEntrada_mapped	<add as new column>	ano_entrada
EffectiveDate	<add as new column>	GETDATE()
AnoLetivo_Inicio_mapped	<add as new column>	LEN((DT_WSTR,8)ano_letivo) != 0 ? SUBSTRING((DT_WSTR,8)ano_letivo,1,4) : ""
AnoLetivo_Fim_mapped	<add as new column>	LEN((DT_WSTR,8)ano_letivo) != 0 ? SUBSTRING((DT_WSTR,8)ano_letivo,5,8) : ""

Figura 57 - Transformação da informação da folha de classificações finais.

De notar que este fluxo de dados apresenta algum tratamento idêntico ao apresentado no processo anterior, nomeadamente na transformação do ano de entrada para quatro dígitos e a validação do ano letivo. Adicionalmente, é feita uma validação semântica sobre as classificações obtidas, sendo que só são consideradas válidas no caso em que a nota numérica seja diferente de zero ou sem valor. Também é validado o número de aluno e a disciplina, dado que são valores cruciais (Figura 58). Todos os registos inválidos encontrados até aqui são armazenados na tabela DQP correspondente.

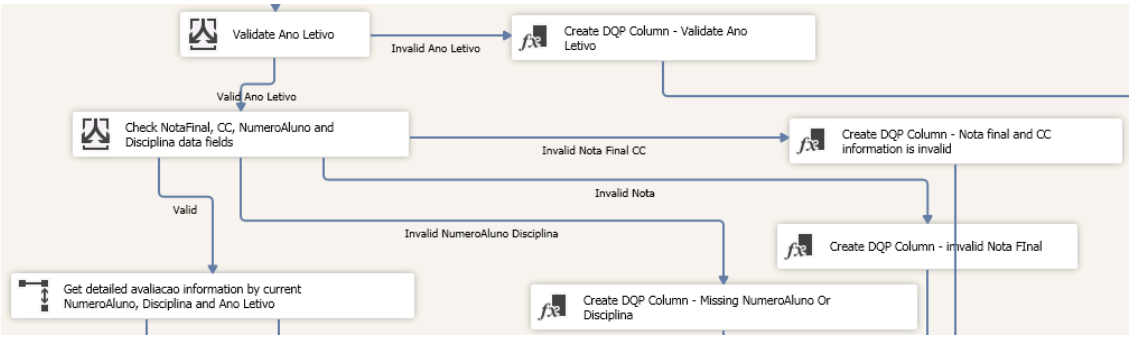


Figura 58 - Validações resultantes da extração e conversão de dados de classificações finais.

Depois de encontrados os registos válidos, o próximo passo passa assim por procurar todos os registos de avaliação existentes carregados na área de *staging*, relativos ao aluno, à disciplina e ao ano letivo/data em questão. No caso de não existirem registos associados, estes são diretamente inseridos na tabela de *staging*. Dado que não existe mapeamento de dados para todos os campos da tabela, optou-se por definir valores por defeito antes da inserção, como demonstra a Figura 59.

Derived Column Name	Derived Column	Expression
Turma_truncated	<add as new column>	-2
Regime_truncated	<add as new column>	"Not Applicable"
Curso	<add as new column>	"LEI"
ClassificacaoExameNorm...	<add as new column>	-2
ClassificacaoExameRecurs...	<add as new column>	-2
ClassificacaoExameEspeci...	<add as new column>	-2
ClassificacaoNotaFinal_tr...	<add as new column>	ISNULL(TipoNota_mapped) == FALSE ? TipoNota_mapped : NotaFinal_truncated >= 9.5 ? "APROV" : "REPRV"
ClassificacaoNotaFrequen...	<add as new column>	-2

Figura 59 – Mapeamento de valores por defeito da fonte de classificações finais.

Caso seja encontrado um registo idêntico, é feita uma verificação da classificação obtida do registo encontrado e do registo corrente, validando se as classificações possuem o mesmo valor de nota. Se forem iguais, é feita uma atualização ao registo dos campos de data e do tipo de classificação, sendo este último atualizado se a nota for obtida por creditação de competências. Se as notas foram diferentes, este registo é considerado inválido, para que possa ser revisto antes da entrada no armazém. A Figura 60 representa como este fluxo está montado.

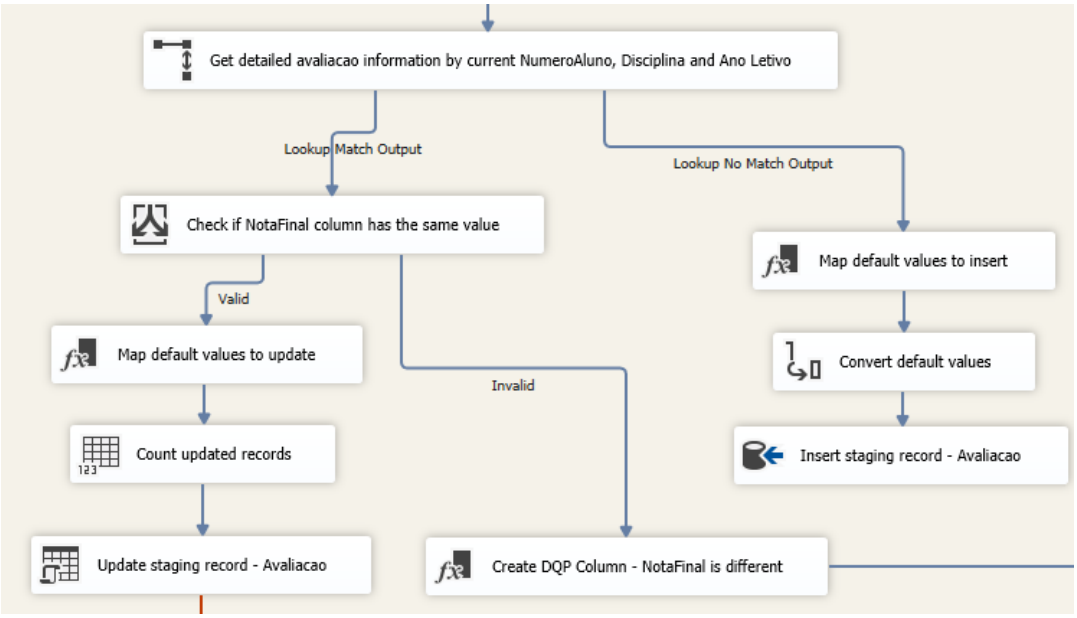


Figura 60 – Fluxo de dados de validação de dados da fonte de classificações finais.

4.2.2 Data Warehouse

Dimensões

Partindo dos dados pré-triados e carregados na área de *staging*, segue-se o processo de carregamento para o AD. Numa primeira fase, foram carregadas as dimensões que obedecem de forma geral ao mesmo tipo de processo. Ao nível do fluxo, a estrutura é composta por um componente de fluxo de dados onde os dados foram extraídos das tabelas, mapeados e inseridos na tabela de destino. Também fazem parte os componentes de auditoria que registam o resultado da execução atual (Figura 61).

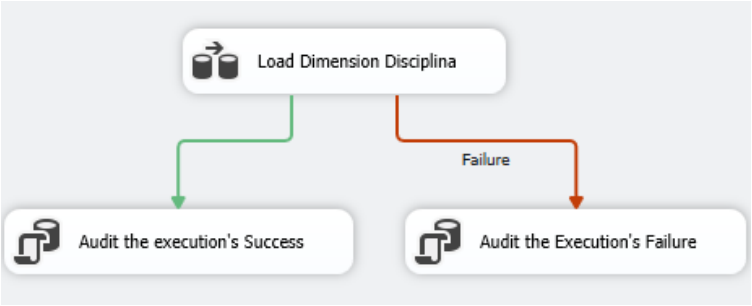


Figura 61 – Componentes do fluxo de carregamento de dimensões.

Dentro do componente relativo ao tratamento de dados, foi feita a ligação com a tabela de *staging* mapeando para o fluxo os campos necessários. De seguida aplicou-se um tipo de manutenção de histórico. Esta técnica (SCD - *Slowly Changing Dimensions*) é utilizada essencialmente com o objetivo de melhor controlar as atualizações aos dados ao longo do

tempo, definindo a necessidade de manter histórico ou não. No desenvolvimento destas tabelas e dependendo das especificidades de cada uma, foi aplicada a técnica de manutenção de histórico do tipo um (SCD1) e/ou a do tipo dois (SCD2).

No caso do primeiro tipo de manutenção de histórico enunciado, é feita uma atualização direta no campo para o valor mais atualizado. Já o segundo tipo é aplicado quando se pretende ter um conjunto de registos onde seja possível identificar as mudanças/atualizações ao longo do tempo, ou seja, sempre que seja feita uma atualização sobre determinado campo, é criada uma cópia desse registo e atualizada a entrada nesse novo registo. Para além da atualização é utilizado um mecanismo para determinar o registo mais recente sendo que, neste caso específico foram utilizadas datas para identificar quando este foi inserido e a partir de que momento é que expirou. Para determinar de forma ainda mais rápida qual o registo mais atual é utilizado um campo booleano, com o valor verdadeiro no registo atual do conjunto de atualizações efetuadas.

Assim, neste contexto optou-se por utilizar os dois tipos de campos para que o momento temporal de expiração seja mais facilmente percetível e para mais facilmente procurar o registo atual para uma determinada chave. A Figura 62 apresenta um exemplo de utilização dos dois tipos de manutenção em simultâneo.

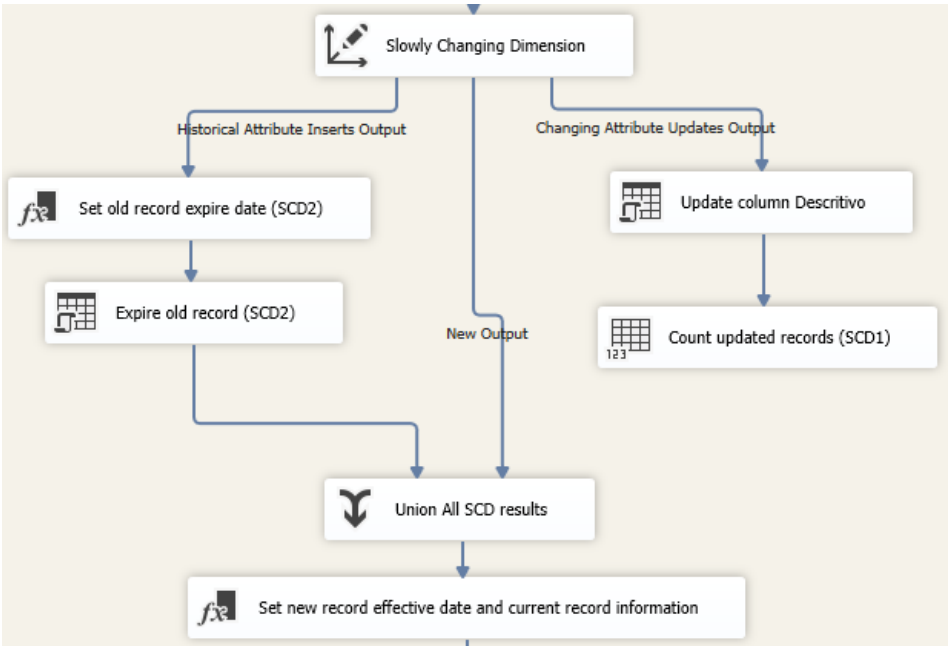


Figura 62 – Exemplo de manutenção de histórico utilizado.

No componente SCD, são definidos para cada campo o tipo de manutenção de histórico pretendido: se o valor é fixo e não se altera, se se pode alterar e o valor é sobreposto ou se é para manter histórico. Podem existir diferentes tipos de manutenção de histórico numa única dimensão, mas aplicados a campos diferentes.

Assim, um registo passa pelo fluxo de histórico representado ("*Historical Attribute Inserts Output*") sempre que o valor de um campo exista para uma determinada chave, criando então o novo registo e mantendo o registo antigo. Nesse registo antigo a variável denominada por "*Expire date*" é atualizada com a data corrente (dado que expirou a partir do momento que o valor necessita de ser atualizado) e o booleano "*IsCurrent*" com o valor falso. O novo registo é inserido com a data de inserção corrente.

O fluxo seguinte ("*New Output"*) não possui qualquer impedimento. Por ele passam todos os registos que o componente SCD considera que não existem, através da comparação com a chave da dimensão. De igual forma ao caso anteriormente descrito, depois de passar todo este fluxo de manutenção, é feita a inserção e contagem do número de registos que vão ser inseridos.

No fluxo que resta ("*Changing Attributes Update Output"*), passam todos os registos que o componente considera que já existem, sendo os campos diretamente atualizados aquando da ocorrência de alterações. São igualmente contados os registos que foram atualizados com sucesso.

Por fim, depois da passagem do fluxo de manutenção de informação, foi feito o mapeamento das variáveis/campos do fluxo com os campos da estrutura para a respetiva inserção de dados.

Como já foi referido, o processo de passagem de dados das tabelas de *staging* para as dimensões apenas difere nos campos e no tipo de manutenção de histórico escolhido. Relativamente ao último ponto, é fulcral que exista a noção temporal de histórico (nos casos possíveis de aplicar) de modo a que todas as análises possam ser efetuadas com dados presentes e passados, sendo assim possível descodificar comportamentos atuais e prevenir futuros. Como pode necessitar de mais recursos disponíveis, capacidade de armazenamento e processamento do armazém de dados, foi escolhido o tipo de manutenção adequado a cada caso, dado que cada tabela é um caso específico, dependendo do tipo de informação e análises que vão ser efetuadas. De seguida apresentam-se as manutenções utilizadas para cada artefacto e respetivas justificações.

- **Dimensão Disciplina:** nesta dimensão, o descritivo da disciplina não se altera com o passar do tempo e, mesmo que altere completamente acaba por se tornar numa edição nova, sendo criado um novo registo de disciplina para associar a uma sigla diferente. Neste caso específico aplicou-se a manutenção de histórico do tipo um para permitir fazer atualização direta no caso de erro (SCD1). Por outro lado, os créditos associados, o semestre, o ano em que foi lecionada podem alterar. Na verdade, não se alteram com frequência, mas considerando que os ECTS têm peso na nota final de um aluno, pode ser interessante analisar o peso que esta tem ao longo das suas edições. O mesmo se passa com a sigla e o semestre, permitindo que alunos que tenham frequentado edições de disciplinas que sofreram alterações sejam na mesma mapeadas. Neste caso aplicou-se uma manutenção de histórico do tipo dois (SCD2);

- **Dimensão Tipo de Classificação:** Não foram consideradas alterações de dados. As siglas de classificação utilizadas são gerais na instituição de ensino a que se destina. Apenas se aplicou uma manutenção de dados do tipo um (SCD1);

- **Dimensão Regime:** Não foram consideradas alterações dado que a informação sobre os regimes que é utilizada é geral a todas as instituições de ensino. Apenas se aplicou uma manutenção de dados do tipo um (SCD1);

- **Dimensão Curso:** Não foram consideradas alterações significativas. Caso os dados do curso se alterem, não é relevante armazenar todas as designações que já teve no passado. Aplicou-se uma manutenção de dados do tipo SCD1.

Tabela de Factos – Avaliações

De uma forma geral, o processo de carregamento de dados para a tabela de factos (Figura 63) é igualmente composto pelo fluxo de carregamento e processo de auditoria. Antes do carregamento começar, é feita uma limpeza à tabela de DQP associada, criada essencialmente com o objetivo de registar possíveis registos problemáticos que não encontrem mapeamento com as dimensões ou até se encontrem duplicados.

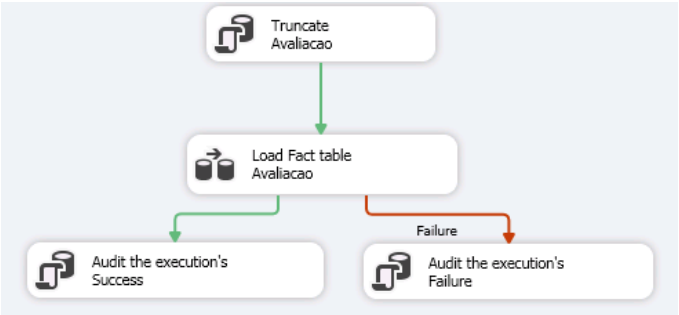


Figura 63 - Componentes do fluxo de carregamento da tabela de factos.

No que toca ao carregamento propriamente dito, foi feito o mapeamento de cada valor para a chave correspondente (Figura 64 representa uma parte de todos os mapeamentos que são efetuados). Para cada coluna existente no fluxo foi aplicada a seguinte lógica:

- **Aluno:** procura exata na tabela de dimensão de alunos por uma chave referente ao número de aluno;

- **Disciplina:** procura exata na tabela de dimensão de disciplinas por uma chave referente à sigla da disciplina;

- **Curso:** procura exata na tabela de dimensão de cursos por uma chave referente à sigla do curso;

- **Regime:** dado que o regime é identificado através de um descritivo, foi feita uma procura na dimensão de regimes referente a uma chave, baseada na semelhança de descritivos. Foi decidido utilizar esta estratégia dado que, na fonte de dados original, o regime é apresentado como um descritivo. Este descritivo pode ter uma construção ligeiramente diferente (utilização de maiúsculas, utilização de diferentes tipos de travessão '-' e '–', entre outros) mas a nível semântico significa o mesmo. Desta forma, foi definido que, determinado regime era nitidamente semelhante, a partir do momento que a sua similaridade seja de 85%;

- **Data:** procura exata na tabela de dimensão de datas por uma chave referente à data de obtenção da classificação (caso exista, tipo numérico no formato yyyyMMdd). Dado que a data de obtenção de classificação não existe na fonte de dados relativa às avaliações dos alunos, esta tem que ser preenchida com um valor por defeito para se conseguir fazer ligação com a data. Nestes casos, passa-se a fazer uma procura por uma chave referente ao ano letivo carregado. Para este efeito, a dimensão data possui uma data por defeito para cada ano letivo diferente.

- **Classificação (notas de frequência, de exames e nota final):** procura exata na tabela de dimensão de classificações por uma chave referente à sigla do tipo de classificação obtido nas notas de frequência, exames (caso existam) ou relativo à nota final.

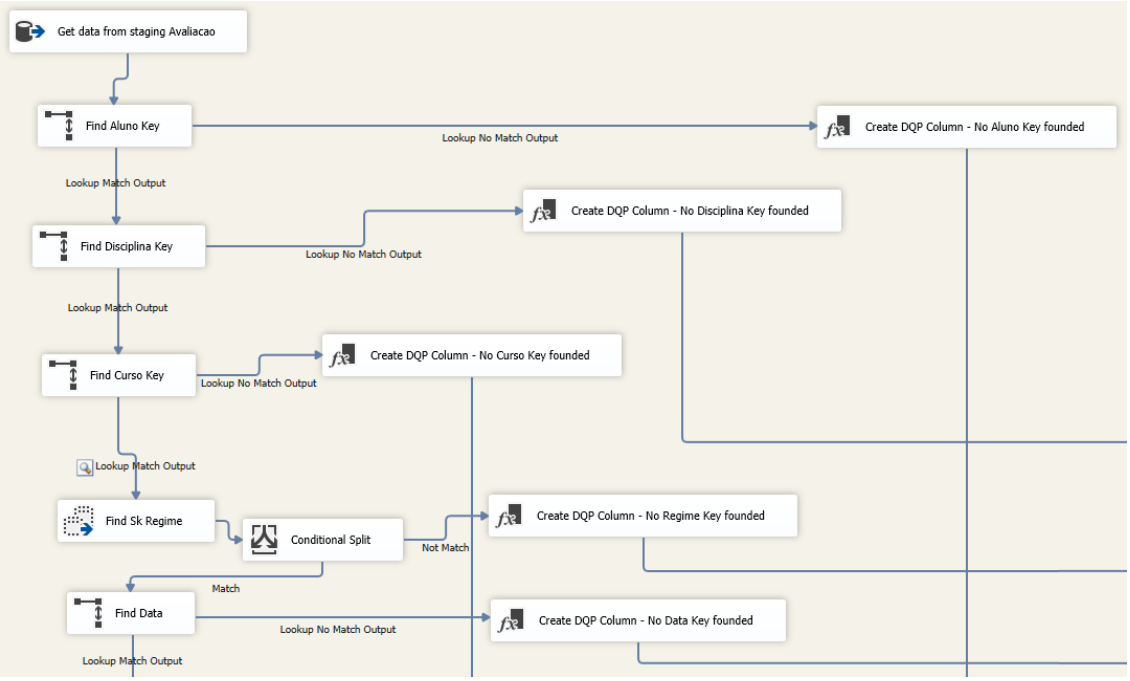


Figura 64 – Extração parcial do carregamento de registos para a tabela de factos.

Antes do armazenamento dos registos, também é verificada a possível existência de duplicados comparativamente aos registos que já existem na tabela de factos. Optou-se por reforçar esta validação apenas para os casos em que os dados são novos e ao qual, erradamente, no meio de

um conjunto de registos estar uma avaliação de um aluno, a uma disciplina e a um determinado ano letivo que já existe.

No caso de sucesso em todas as validações anteriores, os registos são inseridos com sucesso. Caso contrário, quer os registos duplicados quer os registos com dados sem mapeamento são armazenados numa tabela específica para registos inválidos e apenas referente à tabela de factos, permitindo uma análise posterior e decisão/correção do problema. De notar que esta tabela, apesar de conter registos a nível da tabela de factos, está localizada na área de *staging* dado que no AD apenas se devem encontrar registos com dados válidos para serem tidos em conta nas análises de dados.

A Figura 65 apresenta o mapeamento dos dados da tabela de *staging* para armazenamento na tabela de factos de avaliações.

Input Column	Destination Column
Sk_Aluno	Sk_Aluno
Sk_Disciplina	Sk_Disciplina
Sk_Regime	Sk_Regime
Sk_Curso	Sk_Curso
NotaFrequencia	NotaFrequencia
Sk_TipoClassificacaoNotaFrequencia	Sk_ClassificacaoNotaFrequencia
NotaExameNormal	NotaExameNormal
Sk_TipoClassificacaoNotaExameNormal	Sk_ClassificacaoExameNormal
NotaExameRecurso	NotaExameRecurso
Sk_TipoClassificacaoNotaExameRecurso	Sk_ClassificacaoExameRecurso
NotaExameEspecial	NotaExameEspecial
Sk_TipoClassificacaoNotaExameEspecial	Sk_ClassificacaoExameEspecial
NotaFinal	NotaFinal
Sk_TipoClassificacaoNotaFinal	Sk_ClassificacaoNotaFinal
AnoLetivo_Inicio	AnoLetivo_Inicio
AnoLetivo_Fim	AnoLetivo_Fim
Sk_Date	Sk_Date

Figura 65 – Mapeamentos do registo e das variáveis do fluxo para a tabela de factos.

4.3 Cubo de dados OLAP

De uma forma breve, para comprovar a veracidade e a validade dos dados, como primeira abordagem foram replicadas um conjunto de análises utilizando como ferramenta as folhas de cálculo. Para isso, foi criada uma estrutura de dados capaz de agregar toda a informação necessária e que pudesse ser utilizada diretamente pela ferramenta. Assim, na mesma solução onde foi criado o projeto de integração de dados (SSIS), foi adicionado um projeto analítico (SSAS, Figura 66) com o objetivo de criação dessa mesma estrutura denominada por cubo de dados: caracterizando-se como sendo multidimensional, combinada entre dimensões e factos.

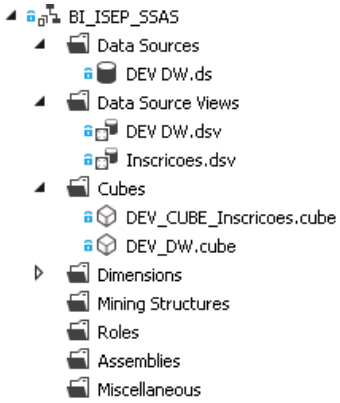


Figura 66 – Estrutura do projeto analítico (SSAS).

Este tipo de estrutura OLAP comporta as medidas definidas na tabela de factos e cada uma das dimensões relacionadas de uma forma otimizada com o objetivo de melhorar os tempos de resposta. Permite que as análises de dados feitas pelos docentes não utilizem diretamente o *data mart*, com o risco de uma possível sobrecarga quando existem operações em paralelo (processo de carregamento *versus* processo de comparar e acumular dados) e torna-se de fácil utilização na medida em que permite realizar várias operações que envolvam cálculos de negócio ao qual no processo de carregamento dos dados pode ser difícil de gerir e/ou manter.

De uma forma breve, a Figura 67 representa toda estrutura lógica de análise criada. Numa primeira fase, foi definida a fonte de dados que aponta para o armazém de dados criado. A partir desta fonte definida, foi criada a respetiva vista que atua como uma camada de abstração, permitindo alterações ao modelo de dados original sem que este sofra qualquer alteração (o que não se aplica neste contexto). Esta vista é composta pelas dimensões e tabelas de factos escolhidas, tendo-se optado por incorporar todos os artefactos existentes. Qualquer alteração feita do lado do armazém de dados requer que esta vista seja atualizada. Todas as ligações são automaticamente replicadas para este modelo, com base nas relações definidas entre tabelas.

Com base nesta vista foi criado o cubo de dados associado, composto pelas dimensões e pelas medidas agrupadas por tabelas de factos. Como já foi referido anteriormente, é de notar que o cubo contém informação que não foi desenvolvida especificamente no contexto desta dissertação (informação sobre alunos, inscrições, entre outras). Optou-se por criar um único cubo que contenha toda a informação, dado que algumas das análises necessitam de alguma informação no âmbito das inscrições.

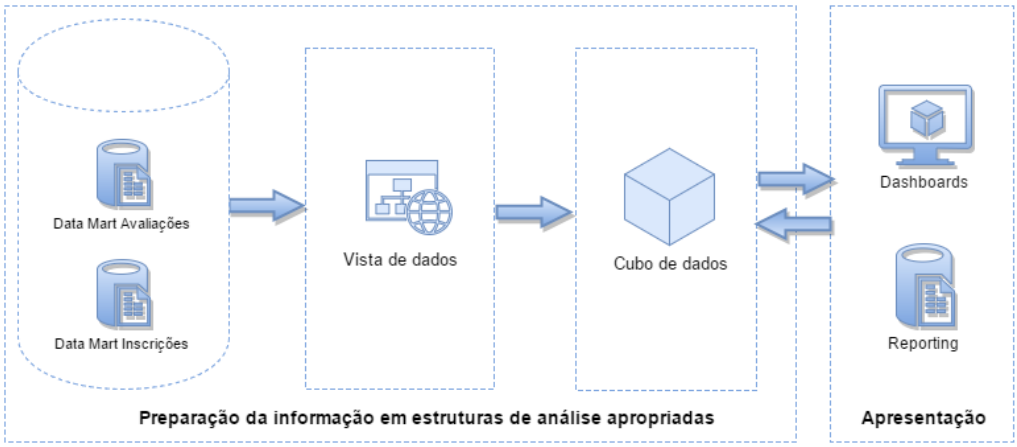


Figura 67 – Estrutura criada para a preparação da camada de apresentação.

4.4 Tabelas agregadas

As estruturas de análise criadas são compostas não só por todas as tabelas de factos e dimensões descritas no contexto de avaliações mas também por tabelas criadas no âmbito das inscrições, dado que algumas das análises requerem informação dos dois contextos.

No que toca à estrutura de avaliações, foram novamente revistas as análises disponibilizadas para tentar compreender o que poderia ser adicionado ao processo para facilitar a replicação das mesmas. Para além das tabelas até aqui apresentadas, encontrou-se a necessidade de criar novas tabelas de factos com um nível de detalhe distinto. Estas tabelas são normalmente utilizadas como técnica de otimização e contém toda a informação carregada na tabela de factos, mas de uma forma agregada, contabilizando alguns cálculos necessários sem a necessidade de re-calcular do início de cada vez que seja feito uma análise sobre a informação. Outra das razões que se prende á criação deste tipo de tabelas deve-se também á quantidade de registos da tabela de factos de avaliações. Neste caso em concreto, as tabelas agregadas contêm menos de metade dos registos que existem na tabela de factos, o que permite também bons tempos de resposta nas consultas.

As tabelas agregadas criadas foram as seguintes (Figura 68):

- **Fact_Avaliacao_Aggreg**: Contabiliza a contagem de inscrições e de tipos de classificações por aluno, ano letivo, ano curricular, regime e horário. É feita uma contagem de aprovações, reprovações, contagem do número de classificações obtidas do tipo sem nota mínima, entre outras;

- **Fact_Avaliacao_Aggreg_Total:** Nesta tabela são consideradas igualmente as inscrições e as contagens de tipos de classificação obtidos pelo aluno, mas a granularidade altera. São calculados os totais, ou seja, as contagens são agregadas apenas por aluno e ano letivo;

- **Fact_Avaliacao_Aggreg_Media:** Esta tabela tem como objetivo agregar informação sobre as notas dos alunos por ano letivo. Para esta listagem são contabilizadas as avaliações cuja nota final obtida seja positiva. É determinada a nota pesada, multiplicando a nota final pelos ECTS da disciplina em questão, são contabilizados o total de ECTS obtidos pelo aluno e é feito o cálculo respetivo da média de notas através da divisão da nota pesada pelo total de ECTS. São igualmente contabilizados o número de anos e o número de inscrições efetuadas;

- **Fact_Avaliacao_Aggreg_Finalista:** Possui o mesmo objetivo anterior. No entanto, apenas agrega informação sobre alunos finalistas, ou seja, que já concluíram ou que estão prestes a concluir o curso. De acordo com a folha de configurações de disciplinas que fazia parte das fontes de dados originais, o número de ECTS para estes alunos terá que ser igual a cento e setenta e nove. É de notar que foi utilizado este valor dado que é o valor que está configurado na folha de dados das disciplinas. O valor real é de cento e ointenta ECTS.

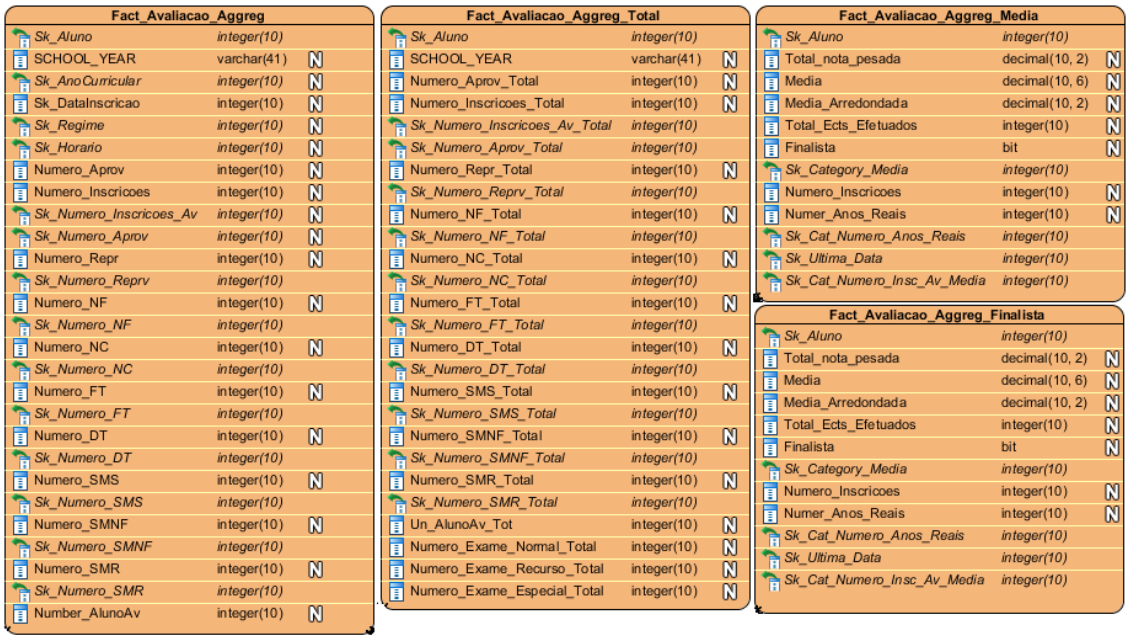


Figura 68 – Tabelas de factos de avaliações agregadas.

É de salientar que todas estas tabelas foram criadas para permitir responder e replicar as análises disponibilizadas. No entanto, nem todas contêm o mesmo nível de informação. Como já foi referido anteriormente, foram disponibilizadas duas fontes de informação distintas: informação detalhada sobre as avaliações dos alunos e informação sobre as classificações finais obtidas. Grande parte das análises apenas contabilizam o primeiro tipo de informação, sendo que a tabela agregada e a agregada de totais apenas contém informação relativa a esse segmento, caso contrário não seria possível obter um termo de comparação. Assim, as únicas tabelas que contêm informação de ambos os contextos são: a tabela agregada de médias e a tabela agregada de finalistas.

Para além das tabelas de factos agregadas, foi criada uma dimensão categoria com o objetivo de auxiliar na identificação das colunas quando se pretende enquadrar a contagem num determinado número de registos, tornando-se possível agrupar os dados por contagens. Tem-se como exemplo a contabilização do número de alunos que esteve inscrito a uma unidade curricular e que obteve aprovação a duas unidades curriculares (ver exemplo em Análise de resultados, a análise sobre disciplinas inscritas versus aprovadas). Resumindo, neste caso esta dimensão permite ter as linhas com a etiqueta “um” e as colunas com etiqueta “dois”.

De notar que a tabela agregada, a agregada de totais e a respetiva dimensão de categorias foram criada em conjunto com o colega Tiago que desenvolveu o *data mart* relativo às inscrições, dado que as análises relacionam os dois contextos. Assim, o cubo de dados também contém as estruturas de dados associadas às inscrições. A Figura 69 representa a estrutura do cubo de dados criado, composto pelas dimensões, tabelas de factos e respetivas relações entre elas. As entidades representadas a amarelo são as tabelas já enunciadas e que não foram desenvolvidas no contexto desta dissertação.

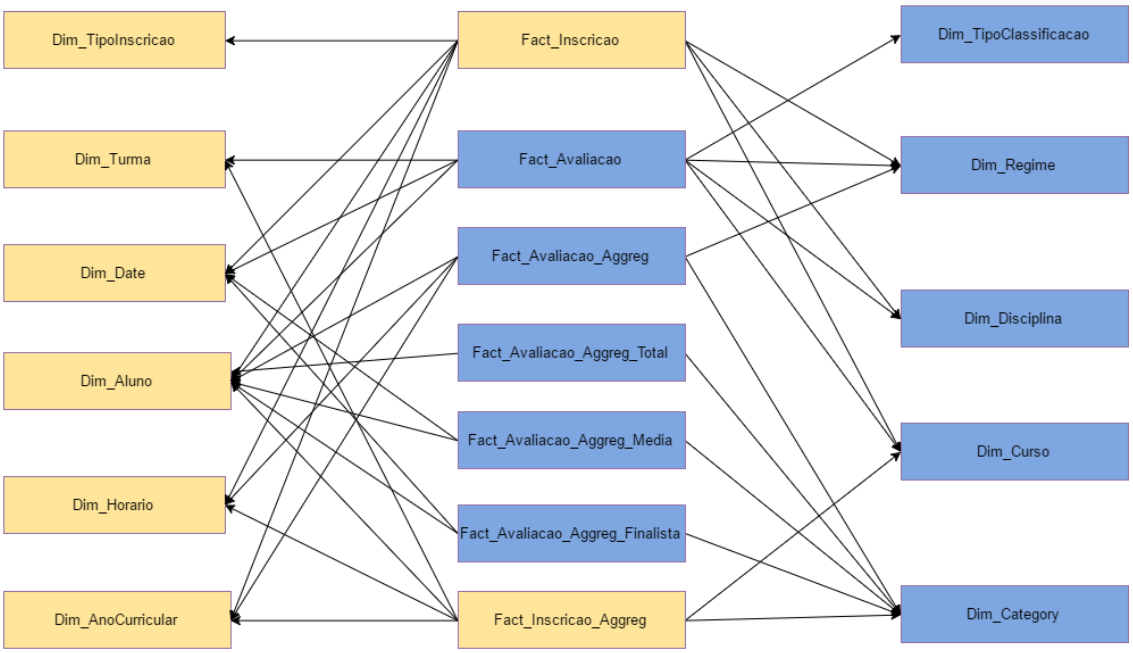


Figura 69 – Estrutura geral do cubo de dados OLAP.

Todas as tabelas agregadas são carregadas com base na tabela de factos. No final do processo de carregamento do AD, foi criado um pacote que executa cada um dos *user stored procedures* criados que alimenta cada uma delas. De uma forma breve, cada procedimento contém operações de soma/contagens e associa cada operação à sua categoria específica.

5 Análise de resultados

Após a fase de desenvolvimento e testes de primeira fase, torna-se importante validar a veracidade do sistema, garantindo que este não altera a validade da informação e que a mantêm coesa. Neste capítulo vai ser apresentada a fase de análise de resultados, refletida essencialmente na realização de análises com a informação tratada e comparada a análises de dados disponibilizadas.

Numa primeira fase são apresentadas um conjunto de análises realizadas, comparadas com o conjunto de análises disponibilizado. As análises realizadas possuem como fonte de dados alvo os dados tratados pelo sistema desenvolvido e ao qual é possível comparar com algumas fornecidas pelo Cliente em dois anos letivos distintos: 2012/2013 e 2013/2014. Estas análises foram replicadas utilizando a mesma ferramenta utilizada nas fontes de análises originais.

Por último, foi utilizada uma ferramenta analítica para mostrar que é possível integrar o sistema de armazenamento desenvolvido com ferramentas analíticas atualmente disponíveis no mercado.

5.1 Comparação de Resultados

Depois da informação tratada e carregada no sistema, foram replicadas algumas análises de diferentes anos letivos (2012/2013, 2013/2014 respetivamente). Dada a extensão necessária para comparação de todas as análises, de seguida vão ser então apresentadas as especificidades e a estrutura de cada uma das análises criadas apenas para o ano letivo 2013/2014. As restantes podem ser consultadas em anexo. O Anexo C contém as restantes análises realizadas sobre o ano letivo corrente em análise. Já no Anexo D e no Anexo E é possível consultar as análises originais disponibilizadas e as análises criadas relativamente ao ano 2012-2013.

Cada uma das análises foi criada com base numa tabela dinâmica (*pivot table*) em *Excel*, direcionada para o cubo de dados. Todas as análises são contempladas com o filtro de dados por ano letivo, o que permite obter os diferentes resultados sem a necessidade de criar análises em separado.

É ainda de salientar que as análises disponibilizadas para termos de comparação sofreram um processo de uniformização dos dados (escolhendo apenas as colunas necessárias para este contexto), de anonimização dado a existência de dados sensíveis relativos aos alunos e uma mudança entre ficheiros de modo a que todos os anos letivos pudessem ser importados a partir de um único ficheiro. Apesar do processo ETL ser revisto várias vezes e de terem sido feitas algumas comparações com as fontes de dados originais, por estes motivos devem-se as diferenças de registos considerável nalgumas análises.

Por outo lado, existem algumas análises sem análises originais de comparação. Foram feitas para melhor comprovar a veracidade da informação carregada e também para demonstrar outro tipo de análises que podem ser feitas sobre os dados.

5.1.1 Unidades curriculares inscritas versus unidades aprovadas

A análise apresentada na Figura 70 tem como objetivo contabilizar o número de alunos, relacionando número de inscrições com o número de aprovações obtidas, para um determinado ano letivo. É feita uma contagem de alunos que se enquadra em cada um dos contextos de unidades inscritas e aprovadas, onde se pode verificar que no total foram contabilizados cento e oitenta e um alunos. Verifica-se também que cento e sessenta e nove alunos inscreveram-se no mínimo a uma unidade curricular e ao qual não conseguiram obter nenhuma aprovação.

A obtenção de aprovações a dez unidades curriculares é a contagem mais elevada de aprovações, contabilizando cento e setenta alunos. O segundo valor mais alto não tem uma diferença significativa, contabilizando cento e sessenta e nove alunos que não obtiveram aprovação a qualquer unidade. Nos valores totais mais altos e anteriores ao referido, verifica-se um valor de aprovações bem menor, que ronda entre as zero e as duas unidades curriculares. No entanto, é possível compreender que muitos poucos alunos estiveram inscritos a doze ou treze unidades e obtiveram aprovação a doze ou mais unidades.

Por outro lado, verifica-se a existência de muitos alunos inscritos a dez unidades curriculares, onde mais ou menos metade obteve aprovação à totalidade. Outro caso que se destaca é o menor caso, sendo que apenas um aluno se inscreveu a catorze unidades.

Número de UCs em que o aluno obteve aprovação																
Nº UCs *	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1	16	28														44
2	12	2	22													36
3	14	6	5	16												41
4	19	5	4	4	8											40
5	26	10	5	12	13	16										82
6	14	9	12	13	14	21	20									103
7	13	7	12	8	12	11	12	16								91
8	12	12	13	12	10	15	10	14	15							113
9	11	7	15	14	12	9	7	7	10	12						104
10	28	16	10	11	19	13	21	28	37	52	150					385
11	4	2	4	2	5	5	7	6	7	9	11	18				80
12		2	3	1	4	5	3	5	1	5	7	3	2			41
13			1			2	1	3	2	2	2	1	2	4		20
14								1								1
15																0
Total	169	106	106	93	97	97	81	80	72	80	170	22	4	4	0	1181

Figura 70 – Análise disponibilizada do número de inscrições versus aprovações.

Com o mesmo tipo de estrutura, foi criada a análise que utiliza então os dados armazenados na solução criada (Figura 71). Tem como fonte de dados a tabela de avaliação agregada de totais, utilizando os campos de contagem de inscrições e aprovações. De notar que foi elaborada em conjunto com o colega Tiago, dado que relaciona informações sobre inscrições e avaliações.

Contagem de Alunos	Número de Aprovações														
Número de Inscrições	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Total Geral
1	20	25													45
2	14	6	18												38
3	12	7	3	16											38
4	20	4	4	5	7										40
5	25	11	5	13	12	15									81
6	15	9	13	15	15	20	17								104
7	13	8	12	10	11	14	11	14							93
8	12	12	13	13	13	13	11	12	13						112
9	11	7	15	13	13	10	7	8	12	11					107
10	27	16	10	11	21	12	20	30	35	54	145				381
11	4	2	4	2	4	5	8	8	6	12	13	10			78
12		2	3	1	4	7	3	4	5	3	7	1	2		42
13			1			2	1	3	2	2	2	2	2	3	20
14								1							1
15	1														1
Total Geral	**174**	**109**	**101**	**99**	**100**	**98**	**78**	**80**	**73**	**82**	**167**	**13**	**4**	**3**	**1181**

Figura 71 – Análise de inscrições e aprovações tendo como base os dados armazenados.

A nível de resultados, é possível verificar que o número de alunos total para o ano letivo em questão é exatamente o mesmo. A nível de contagens verifica-se algumas diferenças, sendo que algumas contagens contabilizam menos registos e outras mais. Os valores máximos para o número de unidades curriculares aprovadas ficou contrário, passando o valor máximo para as zero unidades, seguindo-se logo as dez unidades. O número de aprovações mais baixas manteve-se nas doze ou mais unidades. Foi possível também verificar que o aluno que se tinha inscrito a catorze unidades e foi aprovado a sete manteve-se. No entanto, foi encontrado um aluno que esteve inscrito a quinze unidades, mas não obteve qualquer aprovação. A nível de inscrições, o maior número de inscrições manteve-se nas dez unidades curriculares, com menos quatro registos que a análise anterior. A Figura 72 representa as diferenças entre as análises de forma mais detalhada.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
1	+4	-3													+1
2	+2	+4	-4												+2
3	-2	+1	-2												-3
4	+1	-1		+1	-1										
5	-1	+1		+1	-1	-1									-1
6	+1		+1	+2	+1	-1	-3								+1
7		+1		+2	-1	+3	-1	-2							+2
8				+1	+3	-2	+1	-2	-2						-1
9				-1	+1	+1		+1	+2	-1					+3
10	-1				+2	-1	-1	+2	-2	+2	-5				-4
11					-1		+1	+2	-1	+3	+2	-8			-2
12						+2		-1	+4	-2		-2			+1
13												+1		-1	
14															
15	+1														+1
Total	+5	+3	-5	+6	+3	+1	-3		+1	+2	-3	-9		-1	

Figura 72 – Comparação entre as análises de número de inscrições versus aprovações.

5.1.2 Unidades curriculares inscritas versus unidades com reprovações

Esta análise (Figura 73) possui o mesmo tipo de estrutura e objetivos de análise que a anterior, alterando-se apenas o domínio de informação para reprovações.

	Nº de UCs em atraso							
Nº UCs	0	1	2	3	4	5	6	Total
1	37	7						44
2	27	9						36
3	26	13	2					41
4	17	10	11	2				40
5	39	25	10	3	3	2		82
6	38	31	15	12	5	1	1	103
7	46	10	14	15	2	4		91
8	54	19	12	14	11	2	1	113
9	62	15	13	4	9	1		104
10	338	21	8	11	6	1		385
11	1	44	22	5	2	5	1	80
12			17	17	7			41
13				7	13			20
14					1			1
15								0
Total	685	204	124	90	59	16	3	1181

Figura 73 - Análise disponibilizada do número de inscrições versus reprovações.

Pode-se verificar que o valor mais elevado obtido é de seiscentos e oitenta e cinco alunos que reprovaram a zero cadeiras, o que significa que parte pode ter obtido aprovação ou outro tipo de classificação diferente: não classificado, faltou, entre outras. O valor mais elevado seguinte permite concluir que grande parte dos alunos reprova apenas a uma unidade curricular. Por outro lado, à medida que o número de reprovações aumenta na escala, o número de alunos reduz, sendo que muito poucos alunos reprovam a seis unidades quando inscritos a seis ou mais unidades.

A nível de inscrições verifica-se que o total se mantém constante a nível da análise anterior, dado que se contabiliza os alunos que não obtiveram reprovações, ou seja, que se inserem noutros contextos.

A Figura 74 representa a análise replicada com base nos dados do AD. A fonte de dados é igualmente proveniente da tabela de avaliação agregada de totais, contabilizando os alunos de acordo com o campo de número de inscrições e o número de reprovações. De notar que foi elaborada em conjunto com o colega Tiago, dado que relaciona informações sobre inscrições e avaliações.

Contagem de Alunos	Número de Reprovações						
Número de Inscrições	0	1	2	3	4	5	6	Total Geral
1	41	4						45
2	31	5	2					38
3	32	5	1					38
4	25	11	2	1	1			40
5	48	23	8	1	1			81
6	55	28	17	4				104
7	36	33	13	7	4			93
8	43	33	24	7	4	1		112
9	48	30	11	11	5	1	1	107
10	212	79	39	28	14	6	3	381
11	36	26	8	5	2	1		78
12	8	12	11	10			1	42
13	6	4	7	2	1			20
14			1					1
15			1					1
Total Geral	**621**	**293**	**145**	**76**	**32**	**9**	**5**	**1181**

Figura 74 - Análise de inscrições e reprovações tendo como base os dados armazenados.

Verifica-se que o valor total de contabilizações de alunos é o mesmo. No entanto, para as diferentes contagens entre unidades inscritas e reprovadas existem algumas diferenças significativas. O número de alunos que não obtiveram qualquer reprovação reduziu cerca de sessenta e quatro registos, aumentando o número de reprovações a uma, duas e seis unidades. Todas as restantes contagens por reprovações diminuem com valores menos significativos. A Figura 75 representa detalhadamente as diferenças de contagem de alunos entre as análises.

Nº UCs	0	1	2	3	4	5	6	Total
1	+4	-3						+1
2	+4	-4	+2					+2
3	+6	-8	-1					-3
4	+8	+1	-9	-1	+1			
5	+9	-2	-2	-2	-2	-2		-1
6	+17	-3	+2	-8	-5	-1	-1	+1
7	-10	+23	-1	-8	+2	-4		+2
8	-11	+14	+12	-7	-7	-1	-1	-1
9	-14	+15	-2	+7	-4		+1	+3
10	-126	+58	+31	+17	+8	+5	+3	-4
11	+35	-18	-14			-4	-1	-2
12	+8	+12	-6	-7	-7		+1	+1
13	+6	+4	+7	-5	-12			
14			+1		-1			
15			+1					+1
Total	-64	+89	+21	-14	-27	-7	+2	

Figura 75 - Comparação entre as análises de número de inscrições versus reprovações.

5.1.3 Contagem de exame realizados por ano letivo

A Figura 76 representa uma análise muito simples sobre o número de exames realizados por época e por ano letivo. Dado que apenas estão a ser descritas as análises no âmbito do ano letivo 2013/2014, é possível verificar o maior número de exames de época normal. Este valor alto reflete-se dado que em muitas das unidades curriculares existentes, normalmente existe a obrigatoriedade de exame. De seguida, o valor mais alto é a época de recurso, sendo que foram realizados muito poucos exames especiais.

Ano Letivo	2013-2014	
Total Exames Época Normal	**Total Exames Época de Recurso**	**Total Exames Época Especial**
5908	2052	110

Figura 76 – Análise do número de exames efetuados por ano letivo.

5.1.4 Médias de notas de alunos finalistas

A análise representada na Figura 77 permite determinar o número de finalistas que obteve determinada média final. Foi utilizada a tabela agregada de médias de finalistas, utilizando os campos de categoria relativos ao número de anos de conclusão do curso e média.

É possível verificar que apenas foram encontrados registos de noventa e seis finalistas. Como grande maioria e relativamente à média obtida foram contabilizados mais alunos que obtiveram médias aproximadas a doze, treze e catorze valores. Não foram encontrados valores para médias iguais a dez valores ou superiores a dezassete valores. No entanto, no que toca ao número de anos de demora na conclusão do curso, a grande maioria terminou aos três anos, seguindo-se entre os quatro e os seis anos.

Uma minoria dos alunos demora mais do que seis anos a acabar o curso, nomeadamente oito alunos. Por outro lado, apenas dois alunos conseguiram terminar o curso em dois anos. Isto acontece porque estes alunos em específico obtiveram equivalência (CC) às unidades curriculares do primeiro ano curricular.

Contagem de Alunos	Média de notas						
Número de Anos	11	12	13	14	15	16	Total Geral
2			2				2
3		2	9	14	9	1	35
4		10	11	5			26
5		9	4			2	15
6		9	4				13
7	2	4					6
8	1	1					2
Total Geral	**3**	**35**	**30**	**19**	**9**	**3**	**99**

Figura 77 – Análise de médias de notas de alunos finalistas.

5.1.5 Médias de notas de alunos

Foi criada uma análise com a mesma base que a anterior, mas com o objetivo de contabilizar a média para os alunos no geral (todos), das disciplinas que tenham até ao momento concluídas e cuja nota seja positiva (ou seja, obteve aprovação à unidade curricular). Para este fim foi utilizada a tabela agregada de médias de finalistas. Esta análise é apresentada na Figura 78.

É possível concluir que foram contabilizados trezentos e onze alunos que já possuem unidades curriculares concluídas. Verifica-se que uma grande parte dos alunos possui uma média entre onze e catorze valores, sendo que poucos possuem uma média fora do limite maior referido (cerca de trinta e seis alunos apenas).

Relativamente ao número de anos, é de notar que neste caso não se contabiliza ao número de anos que os alunos demoraram a terminar o curso. Contabiliza-se o número de anos até onde o aluno está no momento, dado que o domínio de informação abrange os finalistas e alunos que ainda estejam a frequentar o curso e ao qual não são finalistas. De uma forma geral, pode-se verificar que o número de anos com contagens mais elevadas são entre os dois anos e os seis/sete anos.

Contagem de Alunos	Média de notas								
Número de Anos	10	11	12	13	14	15	16	17	Total Geral
1	3	5	5	3	4	1	4	2	27
2	7	13	10	11	6	3	1	1	52
3		8	10	20	17	10	1		66
4		5	27	21	5				58
5		7	18	12	2		2		41
6		2	20	7					29
7		9	16	1		1			27
8		2	9						11
Total Geral	**10**	**51**	**115**	**75**	**34**	**15**	**8**	**3**	**311**

Figura 78 – Análise de médias de notas de alunos no geral.

5.2 Resumo dos Problemas de Qualidade de Dados

Foi possível verificar a existência de algumas diferenças relativamente às análises originais. Tal facto se pode dever, como já foi referido, aos processos aos quais as fontes de dados foram alvo e também ao facto de existir informação inconclusiva/inválida.

Para melhor compreender o sucedido, as tabelas de DQP armazenaram ao longo do processo todos os registos considerados como inválidos. A Tabela 9 representa de uma forma resumida, os problemas encontrados no carregamento dos três anos letivos enunciados relativamente à folha de notas de alunos e à folha de fichas de aluno (classificações finais).

Tabela 9 – Análise de registos inválidos na fonte de dados de Notas de Alunos.

Geral - Ano Letivo 2013-2014	
Motivo	**Quantidade de registos**
Registos com a época inválida (“C2”,”DZ”,”PO”)	365
Registos duplicados	21
Registos em branco (sem aluno, disciplina, regime, curso, época, ano letivo e notas)	4
Registos onde a nota final negativa	11
Registos encontrados durante o processamento que, ao serem comparados com os dados das notas de aluno inseridos na tabela de *staging*, foram encontrados registos para o mesmo aluno, para a mesma disciplina e para o mesmo ano letivo. No entanto a nota final não corresponde à que já foi inserida.	206

Relativamente à primeira folha, verificou-se a existência de alguns registos com épocas inválidas. No desenvolvimento da solução foi definido um conjunto de épocas sendo apenas estes considerados como válidos. Neste caso, as épocas tomam os valores “C2”,”DZ” e ”PO” (Figura 79), aos quais não existem na dimensão de regimes.

Numero	Nome	NFreq.	Exame	Nota	Disciplina	Época
1000051			11	10	ALGAV	C2
1000051			1,8	7	FSIAP	C2
1010066			5,8	6	IARTI	C2
1010069			FT	NC	IARTI	C2
1020082			FT	SMS	SGRAI	DZ
1030083			16	16	PESTI	C2
1030093			15	15	PESTI	C2
1040101			12	12	PESTI	DZ
1040109			11	11	PESTI	DZ
1040111			FT	NC	BDDAD	C2
1040111			11,2	13	EAPLI	C2
1040111			FT	NC	ESINF	C2

Figura 79 – Extração da folha de notas de alunos - regimes inválidos.

Na Figura 80 é apresentado um exemplo dos registos que foram encontrados como duplicados a nível do número de aluno, disciplina, época e ano letivo.

Numero	Nome	NFreq.	Exame	Nota	Disciplina	Época
1030093			FT	NC	PESTI	NM
1030093			FT	NC	PESTI	NM

Figura 80 – Extração da folha de notas de alunos - registos duplicados.

Por último, foram encontrados alguns registos com números de aluno ao qual não existiam na dimensão Aluno. Estes alunos não foram carregados para essa dimensão devido ao facto de os registos associados não possuírem pelo menos uma inscrição a uma turma/disciplina. A Figura 81 apresenta o exemplo de um aluno sem especificação da inscrição.

Número	Nome	Regime Freq.	Ano Curr	D/N	ALGAN	APROG	AMATA	LAPR1	PRCMP	ARQCP	BDDAD	ESINF	FSIAP	LAPR3	ASIST	ALGAV	ARQSI	GESTA	LAPR5	SGRAI	ESOFT	LAPR2	MATCP	MDISC	PPROG	EAPLI	LAPR4	LPROG	RCOMP	SCOMP	CORGA	COMPA	IARTI	PESTI	EST-ERASMUS 10	EST-ERASMUS 20	EST-ERASMUS 30
1121226		Vasco da G	3 (n/i)																																			

Figura 81 - Extração da folha de notas de alunos – alunos sem mapeamento com a dimensão aluno.

No caso da segunda fonte de dados (folha de classificações finais), foram encontrados apenas quatro registos em branco, sem qualquer tipo de informação. Não é visível na folha de cálculo, no entanto uma causa para este problema poderá ser quando se limpa os dados de células, sendo que o componente de extração considera que aquela célula teve qualquer tipo de edição, apesar de não existir nenhum conteúdo textual. De seguida, foram detetados alguns registos ao qual a nota final é negativa (Figura 82). Considerando o tipo de nota (TN), não existe a informação sobre qual foi a nota obtida por equivalência não superior. Optou-se por não carregar essa informação nesta fase.

Número	Disciplina	Nota	Data	TN	AnoEnt
1090518	RCOMP	0	19/09/2013	FNS	09
1100657	GESTA	0	19/09/2013	FNS	10
1110747	APROG	0	01/10/2014	FNS	11
1110747	PRCMP	0	01/10/2014	FNS	11
1110747	RCOMP	0	01/10/2014	FNS	11
1110913	PRCMP	0	23/09/2013	FNS	11
1110913	RCOMP	0	23/09/2013	FNS	11
1110942	APROG	0	27/11/2014	FNS	11
1110942	PRCMP	0	27/11/2014	FNS	11

Figura 82 - Extração da folha de classificações finais - notas finais negativas.

Por fim, a Figura 83 representa um conjunto de registos que, quando comparados com as notas finais do registo correspondente (para o aluno, disciplina, ano letivo) da nota de alunos, os valores diferiram. A primeira tabela representa a fonte de dados de notas de alunos, onde a nota final obtida foi de dezasseis valores. No entanto, a segunda fonte indica que a nota final obtida foi de dezassete valores.

Numero	Nome	NFreq.	Exame	Nota	Disciplina	Época
1030083		18	14	16	ARQSI	NM

Número	Disciplina	Nota	Data
1030083	ARQSI	17	10/02/2014

Figura 83 – Extração de cada uma das fontes de dados onde as notas finais não coincidem.

5.3 Integração com o *Power BI Desktop*

Para melhor comprovar a utilização da solução implementada por parte de outras ferramentas analíticas, foi utilizado o *software Power BI Desktop* para reproduzir algumas análises. Numa primeira fase, foi criado um projeto cuja fonte de dados aponta para o cubo de dados OLAP criado. A aplicação carregou todos os artefactos nele existentes (tabelas de factos, tabelas de factos agregadas e respetivas dimensões) sendo a criação de gráficos no *dashboard* muito simplificada, escolhendo numa primeira abordagem o tipo de gráfico que se pretende e definindo o que se pretende mostrar no mesmo fazendo *drag-and-drop* dos campos necessários para a parte dos valores do gráfico, filtros e valores de eixos. Um *dashboard* pode ser composto por um ou mais gráficos de informação, sendo que cada um pode ser independentemente personalizável também a nível de aspeto como cores, títulos, unidades de medida, entre outros parâmetros.

A Figura 84 representa um *dashboard* criado relativamente a informação do ano letivo 2013-2014, que contém três análises já anteriormente descritas em diferentes tipos de gráficos. No gráfico de barras é feita a análise dos alunos aprovados por ano de entrada e ano curricular, no gráfico circular é contabilizada a contagem de exames realizados por época e, por fim, o número de reprovações por ano curricular. É de salientar que na análise dos aprovados/reprovados não foram contabilizados os alunos que obtiveram zero aprovações/reprovações.

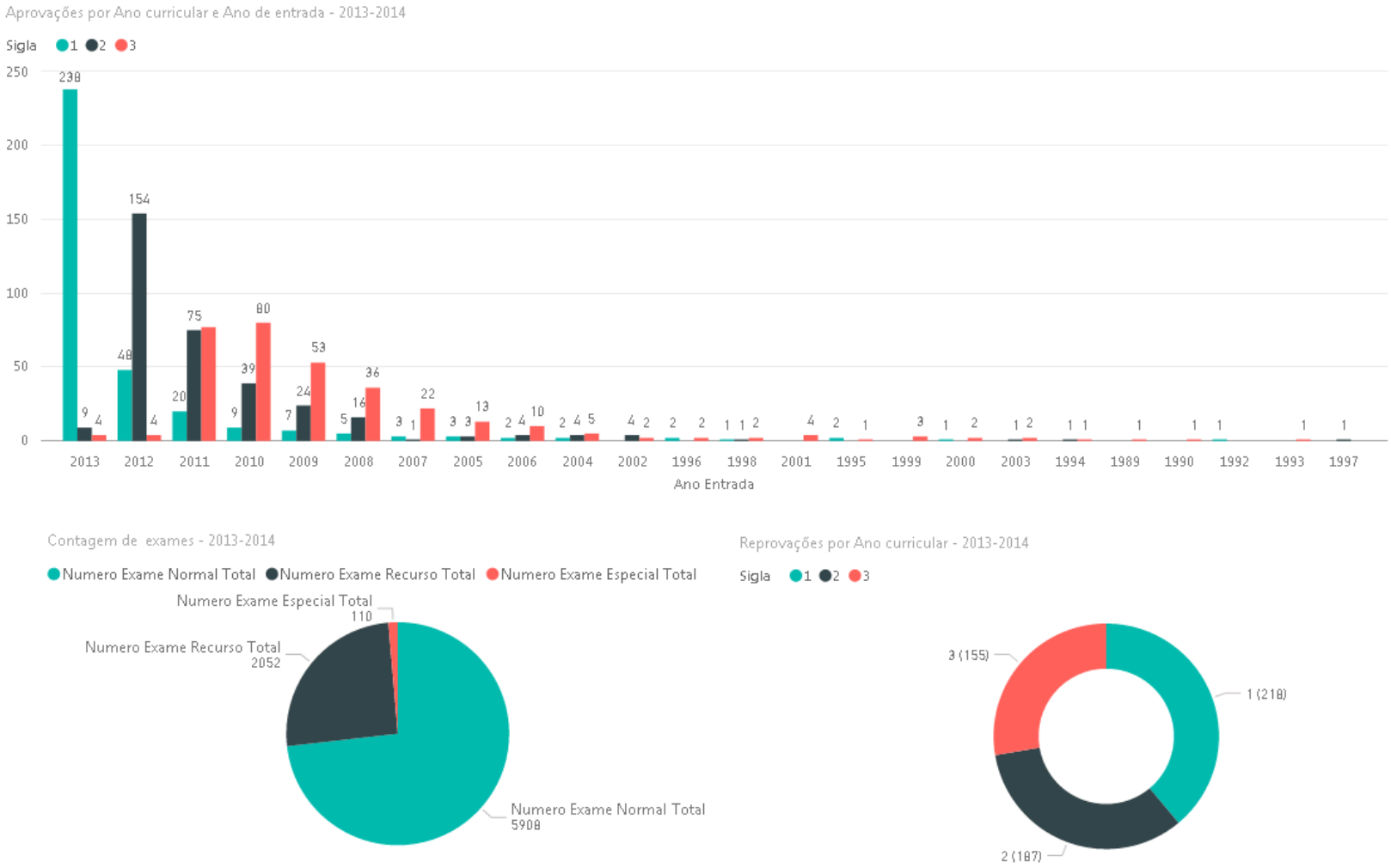


Figura 84 – Integração com o *Power BI Desktop*.

6 Conclusão

Cada vez mais as instituições de ensino apostam no uso das tecnologias de informação para o apoio e monitorização das tarefas de aprendizagem. Para além de tornarem o processo de tratamento de dados mais fácil e controlado, estas fornecem um conjunto alargado de ferramentas com o objetivo de servir para levantamento de comportamentos e o suporte de decisões. Neste sentido, o Diretor da LEI sentiu a necessidade de tratar a informação sobre as avaliações dos alunos com o objetivo de monitorizar todos os acontecimentos e a produtividade escolar.

Com os avanços tecnológicos, aparecem cada vez mais ferramentas que permitem desenvolver uma solução capaz de responder às necessidades do meio. No entanto, o cliente optou pelo desenvolvimento da sua própria solução não só por questões monetárias, mas também para, numa primeira fase, ser desenvolvido um protótipo que o ajude nas suas tarefas de avaliação de resultados do ano escolar.

6.1 Objetivos alcançados

Nesta dissertação foi possível apresentar uma solução cumprindo todos os objetivos iniciais. De forma mais explícita, esta solução desenvolvida para responder às necessidades específicas da LEI que permite limpar, transformar e carregar a informação disponibilizada sob a forma de folhas de cálculo para um repositório de dados centralizado e específico para o ambiente escolar e para manutenção de informação de avaliações de alunos. Esta lógica foi definida com base na análise da estrutura/tipo de dados existente nas fontes e no conjunto de análises definidas como importantes para o cliente final. Para comprovar o seu funcionamento foram feitos vários carregamentos para o repositório de dados, com o objetivo aperfeiçoar o processo, corrigindo eventuais problemas no carregamento e no tratamento da informação.

Por outro lado foi criado também um conjunto de análises com o objetivo de serem comparadas as análises originais com as criadas. Optou-se por reproduzir as mesmas na mesma ferramenta utilizada para as análises originais (Excel).

Apesar de não ser um objetivo explícito, também foi possível demonstrar uma possível integração da solução criada com uma ferramenta de apresentação de dados do mercado e diferente das folhas de cálculo utilizadas, permitindo fazer variadíssimas consultas alterando apenas as variáveis necessárias. Desta forma, o cliente tem agora a possibilidade de integrar este repositório de dados com outras ferramentas analíticas, podendo criar um conjunto de relatórios de informação de forma mais coesa, para analisar dados dos alunos dos diferentes anos letivos.

Uma das grandes vantagens que esta solução traz para os possíveis utilizadores é sem dúvida o tempo gasto no tratamento, na realização de cálculos e criação de relatórios/análise de dados. Apesar das fontes de dados terem que ser extraídas e armazenadas em ficheiros (ao qual pode resultar em tempo burocrático significativo), os passos necessários à extração e tratamento dos dados passam a ser automáticos e com menor probabilidade de erros de semântica, sendo o tempo de transformação considerável a partir do momento que existe um número elevado de registos. No entanto, as análises de dados passam a ter como base um repositório de dados centralizado e mais fiável, com a flexibilidade de realização de análises parametrizadas por qualquer tipo de dados existentes no AD. Como foi dito anteriormente, dado que é possível integrar o AD com ferramentas que permitam criar *dashboards* com uma organização da informação mais familiar para o utilizador final, o tempo de criação dessas vistas e respetiva análise aos dados pode reduzir em massa dado que não são necessários conhecimentos técnicos ao nível das tecnologias de informação para criar e analisar esses *dashboards*

6.2 Trabalho futuro

Apesar do AD desenvolvido ser uma possível solução para a resolução do problema apresentado, tal como em todos os produtos existem aspetos que podem ser melhorados.

A nível do processamento da informação, destaca-se a possibilidade de definir uma nota mínima por unidade curricular. No futuro, permitiria elaborar cálculos específicos de notas para cada disciplina em cada uma das suas componentes. Esta informação não foi facultada para o carregamento do sistema nesta fase, contudo este é adaptável a esta alteração: capacitando a dimensão disciplina com mais um atributo e alterando este tipo de validações no fluxo de dados.

Outro aspeto que poderia ser alvo de melhoria seria a deteção de duplicados no carregamento de dados. Neste momento é feita a procura de duplicados no fluxo de carregamento das tabelas de *staging* de avaliações, identificando cada linha prestes a ser carregada com um valor inteiro. Com o objetivo de não existirem impactos ainda mais elevados a nível das análises comparativas, todas as linhas com índice um, são carregadas, sendo que todas as outras que possuem um índice diferente são detetadas como duplicadas e inseridas na tabela DQP respetiva. A melhoria

neste ponto passaria por passar a responsabilidade para o humano de definir qual registo do conjunto que é considerado como unívoco/válido e que entra para o AD, ou seja, nenhum registo passaria para o AD a partir do momento em que fosse encontrado um registo duplicado. Todos eles iriam ser armazenados na DQP.

Seria também interessante a existência de um processo automático de gestão e importação de ficheiros. Este adicional passaria por criar uma aplicação associada à solução apresentada que permitiria escolher o ficheiro de dados ou conjunto de ficheiros que eram necessários importar, gravar a listagem de ficheiros que já foram carregados para o AD e visualização gráfica dos resultados de execução de cada carregamento. Não foi feito nesta fase dado que a periodicidade a que o carregamento é feito é apenas de ano a ano, ou seja, quando termina cada ano letivo. Seria importante apenas numa primeira fase para lançar de uma só vez os dados de um conjunto de anos letivos.

Outro aspeto importante seria analisar a performance da solução ao longo do tempo e nos diferentes carregamentos. Neste caso em específico apenas foi disponibilizado um único ficheiro de dados para carregamento, pelo que não era possível este tipo de análises. No entanto, esta análise passaria pela comparação de tempos de processamento e de utilização de recursos computacionais perante as capacidades de *hardware* existentes.

Por fim, e abstraindo do contexto técnico, dado que as tecnologias estão sempre a evoluir e com isto, novas ferramentas de desenvolvimento e novas formas de implementação surgem, pode-se assim aferir que o AD desenvolvido não é um projeto concluído. Apesar de se apresentar como protótipo, este AD permite ser adaptado ao longo do tempo. Também os processos escolares podem sofrer alterações no futuro, fazendo sentido criar novas áreas de negócio ao qual seja do interesse extrair conhecimento como informação sobre novas licenciaturas, mestrados e até informação de outros departamentos existentes na instituição. Aliado aos novos dados, também podem surgir novos KPI's com interesse para análise.

Referências

1keydata 2015, *MOLAP, ROLAP, and HOLAP*, viewed 20 September 2015, <http://www.1keydata.com/datawarehousing/molap-rolap.html>.

ARSON Group SAC 2016, 'Oracle Data Integrator', *ARSON Group SAC*, viewed 22 June 2016, <http://www.arsongroup.com/web/productos/oracle-data-integrator/>.

Atanazio, J. 2013, 'Conceituando BI – Parte IV: Diferenças entre OLTP x OLAP', *bisaopaulo.com*, viewed 20 September 2015, <http://bisaopaulo.com/bi/conceituando-bi-parte-iv-diferencas-entre-oltp-x-olap/>.

Ballard, C., Farrell, D.M., Gupta, A., Mazuela, C. & Stanislav, V. 2006, *Dimensional Modeling: In a Business Intelligence Environment*, First Edition.

Branislav Barnak, Amir Bar-or, Cynthia M.Saracco & Paul Stanley 2009, 'IBM InfoSphere DataStage e DB2 pureXML, Parte 2: Desenvolvendo um Armazém de Dados Ativado para XML', *IBM developerWorks*, CTZZZ, viewed 25 June 2016, <http://www.ibm.com/developerworks/br/data/library/techarticle/dm-0909datastagepurexml2/sidefile-fig1.html>.

CaseNex 2010, *DataCation*, viewed 31 January 2016, <http://www.datacation.com/Services/Data%20Warehousing/>.

datawarehouse4u 2009, *OLTP vs. OLAP*, viewed 16 September 2015, <http://datawarehouse4u.info/OLTP-vs-OLAP.html>.

Dean, T. 2015, *Gain the Competitive Edge with Business Intelligence Software & Analytics*, viewed 15 September 2015, <http://www.business2community.com/business-intelligence/gain-competitive-edge-business-intelligence-software-analytics-01280045>.

Editorial Team+ 2007, 'What is Operational Database | Online Learning', viewed 16 September 2015, <http://www.learn.geekinterview.com/data-warehouse/dw-basics/what-is-operational-database.html>.

Elias, D. 2015, 'Entendendo a modelagem multidimensional - Business Intelligence', *Canaltech*, viewed 15 September 2015, <http://corporate.canaltech.com.br/materia/business-intelligence/entendendo-a-modelagem-multidimensional-19988/>.

ETL, D. 2015, 'ETL Tools - Top 10 ETL Tools Reviews - Database ETL', viewed 22 September 2015, <http://www.databaseetl.com/etl-tools-top-10-etl-tools-reviews/>.

'Etl tools comparison' 2016, viewed 13 February 2016, <http://www.etltools.net/etl-tools-comparison.html>.

FenProf 2012, 'O sistema de ensino Superior em Portugal'.

Ferreira, E. 2015, 'Experimentação e avaliação'.

G2 Crowd 2016, 'MicroStrategy vs. Power BI | G2 Crowd', *Compare MicroStrategy vs. Power BI*, viewed 13 February 2016, <https://www.g2crowd.com/compare/microstrategy-vs-microsoft-power-bi>.

IBM 1999, 'DATA WAREHOUSE OPERATIONAL ARCHITECTURE', p. 14.

IBM 2015, 'IBM Knowledge Center - Business Monitor KPIs', *Key performance indicators (KPIs)*, viewed 21 June 2016, <http://www.ibm.com/support/knowledgecenter/SS7NQD_8.0.0/com.ibm.wbpm.wid.tkit.doc/model/kpis.html>.

IBM 2006, 'School District of Philadelphia improves performance with data-driven decision making', p. 4.

InformationWeek 2015, 'Put to the Test: Oracle Data Integrator - InformationWeek', *Put to the Test: Oracle Data Integrator - InformationWeek*, viewed 23 December 2015, <http://www.informationweek.com/software/information-management/put-to-the-test-oracle-data-integrator/d/d-id/1054359?page_number=1>.

Inmon, W.H. 2002, *Building the Data Warehouse*, 3º., John Wiley & Sons, Inc., Canada.

Jet Reports 2016, 'Junk Dimension Automation with the Jet Data Manager', *Jet Reports Knowledge Base and Resources*, viewed 31 August 2016, <http://jetsupport.jetreports.com/hc/en-us/articles/218952368-Junk-Dimension-Automation-with-the-Jet-Data-Manager>.

Kimball, R. & Caserta, J. 2004, *The Data Warehouse ETL Toolkit*, Wiley Publishing, Inc., Canada.

Kimball, R. & Ross, M. 2013, *The Data Warehouse Toolkit Third Edition*, 3ª., John Wiley & Sons, Inc., Canada.

Luhn, H.P. 1958, 'Luhn, H. P., 1958. A Business Intelligence System. IBM JOURNAL, Issue A Business Intelligence System, p. 6.', *IBM JOURNAL*, p. 6.

McBurney, V. 2007, *Wiki Wednesday: comparing Talend and Pentaho Kettle open source ETL tools*, viewed 19 February 2016, <http://it.toolbox.com/blogs/infosphere/wiki-wednesday-comparing-talend-and-pentaho-kettle-open-source-etl-tools-16294>.

Microsoft 2016a, *Power BI - Descrição Geral e Aprendizagem - Power BI*, viewed 13 February 2016, <https://support.office.com/pt-pt/article/Power-BI-Descri%C3%A7%C3%A3o-Geral-e-Aprendizagem-02730e00-5c8c-4fe4-9d77-46b955b71467>.

Microsoft 2016b, *Sobre cubos OLAP*, viewed 24 August 2016, <https://technet.microsoft.com/pt-br/library/hh916536(v=sc.12).aspx>.

Microsoft 2015, *SQL Server Integration Services*, viewed 26 December 2015, <https://msdn.microsoft.com/en-us/library/ms141026(v=sql.120).aspx>.

Microsoft 2010, 'Synchronize two tables using SQL Server Integration Services (SSIS)–Part I of II', *SQL Server Cast*, viewed 27 August 2016, <https://blogs.msdn.microsoft.com/jorgepc/2010/12/07/synchronize-two-tables-using-sql-server-integration-services-ssispart-i-of-ii/>.

MicroStrategy 2015, 'About MicroStrategy | MicroStrategy', *Soluções Business Intelligence, Analytics & Mobile | MicroStrategy*, viewed 13 February 2016, <https://www.microstrategy.com/pt/sobre-nos/sobre-nos>.

Narasimharajan], M. 2011, 'Integration, Maximize Success on Data'.

New York University 2014, *University Data Warehouse Plus*, viewed 2 February 2016, <https://www.nyu.edu/employees/resources-and-services/administrative-services/university-data-warehouse-plus.html>.

Nubera 2015, 'Phocas Pricing, Features, Reviews & Comparison of Alternatives', *GetApp*, viewed 21 February 2016, <https://www.getapp.com/business-intelligence-analytics-software/a/phocas/>.

OLAP.com 2016, *What is the Definition of OLAP? OLAP Definition*, viewed 20 September 2015, <http://olap.com/olap-definition/>.

Oracle 2007, *Data Mart Concepts*, viewed 17 September 2015, <http://docs.oracle.com/html/E10312_01/dm_concepts.htm>.

Oracle 2002, *Data Warehousing Concepts*, viewed 24 September 2015, <http://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm>.

Oracle, O. 2016, 'Oracle Data Integrator', *Oracle Data Integrator*, viewed 13 February 2016, <http://www.oracle.com/technetwork/middleware/data-integrator/overview/index.html>.

Patusky, C., Botwinik, L. & Shelley, M. 2007, *The Philadelphia SchoolStat Model*, Philadelphia.

Pentaho 2015, *Data Integration | Pentaho Community*, viewed 29 December 2015, <http://community.pentaho.com/projects/data-integration/>.

Pereira, N. 2016, 'Escrita Técnico - científica'.

Phocas, S. 2015, *Phocas: Successful Business Intelligence Software and Data Discovery*, viewed 19 February 2016, <http://www.phocassoftware.com/>.

Rouse, M. 2015, 'Mobile BI tools, trends and best practices guide', *TechTarget*, viewed 15 September 2015, <http://searchbusinessanalytics.techtarget.com/essentialguide/Mobile-BI-tools-trends-and-best-practices-guide>.

Sansu George 2012, 'Inmon vs. Kimball: Which approach is suitable for your data warehouse?', *ComputerWeekly*, viewed 25 June 2016, <http://www.computerweekly.com/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse>.

Saúde, S., Borralho, C., Féria, I. & Lopes, S. 2014, 'A necessária especificidade da avaliação de desempenho das Instituições de Ensino'.

School Buffalo 2015, *EdVantage Data Dashboard*, viewed 31 January 2016, <http://www.buffaloschools.org/informationtech.cfm?subpage=66927>.

SchoolCity Inc. 2015, *School Data Warehousing | Edvantage™*, viewed 31 January 2016, <http://www.schoolcity.com/data-warehousing.html>.

Tech-FAQ 2013, *Data Warehouse*, viewed 16 September 2015, <http://www.tech-faq.com/data-warehouse.html>.

tutorialspoint 2016, 'Data Warehousing OLAP', *www.tutorialspoint.com*, Data Warehousing - OLAP, viewed 25 September 2016, <https://www.tutorialspoint.com/dwh/dwh_olap.htm>.

Anexo A

Análise de valor

A análise de valor pode ser considerada como uma metodologia utilizada para compreender e descobrir o rumo certo para a criação de valor no mercado. Baseia-se na identificação das necessidades não satisfeitas, permitindo reduzir custos na produção sem sacrificar a qualidade, utilizando apenas os recursos necessários e com foco no que é essencial para satisfazer as necessidades encontradas. É igualmente importante compreender como o mercado está estruturado e composto, identificando também os possíveis concorrentes e como marcam ou podem vir a marcar a diferença. Desta forma, a estratégia de atingir os consumidores alvo pode ser mais facilmente compreendida e alterada, com o objetivo de oferecer mais qualidade e inovação no produto.

Como foi descrito, o valor é um fator muito importante a ter em conta nesta análise e pode ser definido pela importância que se confere a determinado bem. Na perspetiva do cliente, este atribui valor a um determinado produto não só de acordo com as necessidades que possui, tendo em conta a quantidade de bens disponíveis para as satisfazer, mas também nos benefícios e sacrifícios em causa (*Value to Customer*). Do lado do fornecedor o importante é distribuir, manter e melhorar o produto progressivamente, possibilitando a angariação de novos consumidores e a manutenção dos atuais. A estas descrições de valor dá-se o nome de valor percebido, englobando uma definição de valor visto de duas perspetivas diferentes.

Este ainda pode assumir diferentes formas: através do balanço dos benefícios e sacrifícios de forma a disponibilizar o melhor ou o pior valor para o consumidor (*Net Value to Customer*), através dos atributos percebidos (*Marketing Value to Customer*), valor definido apenas pelo seu preço (*Sale Value to Customer*), valor definido apenas pelo seu preço objetivo/sentimentalista (*Rational Value to Customer*) e, por fim, ainda pode ser definido através do *feedback* das experiências de consumo dos clientes (*Derived Value to Customer*).

Neste contexto, foi elaborado um modelo Canvas (apresentado na secção Modelo de Canvas) com o objetivo então de definir a criação de valor e os sacrifícios que esta solução traz aos intervenientes.

A solução apresentada neste documento irá ser desenvolvida por alunos do MEI a partir de servidores disponibilizados pelo Departamento, criando *know-how* sobre a informação curricular a partir das necessidades apresentadas, sendo *a posteriori* utilizada pelo Diretor da LEI (Cliente Direto/Final). Estes são os elementos chave deste processo, não excluindo a possibilidade de este produto vir a ser divulgado pelos restantes docentes (inclusive pelo Diretor do MEI) e também pelos restantes Departamentos da instituição (possíveis canais de divulgação). Pode-se assim concluir que ambas as partes saem a ganhar (modelo *win-win*), na

medida em que os produtores ganham experiência/reconhecimento e o Cliente final obtém uma solução à medida das suas necessidades.

Assim, como proposta de valor e depois de desenvolvida esta solução, a informação escolar passa a estar centralizada e armazenada num armazém de dados especializado, acessível por parte dos docentes responsáveis pelo departamento (nomeadamente o Diretor da LEI) e apto para suporte às análises de dados necessárias. Dado que é uma solução desenhada à medida das necessidades do Departamento, o acesso é efetuado de forma rápida e auxilia em diversos cálculos matemáticos, permitindo assim reduzir custos no tempo necessário, tanto de compreensão e tratamento de erros, bem como no tempo de extração de informação.

Para além do desenvolvimento da solução, como atividades chave será necessário manter os servidores utilizados para que o risco de indisponibilidade de dar resposta seja reduzido. Esta manutenção de servidores terá que ser feita com a colaboração dos Técnicos, responsáveis por todo o equipamento informático/tecnológico do Departamento. Será também necessário dar assistência técnica, suporte ao que foi desenvolvido a nível do AD, no processo de carregamento de dados e ainda suporte na criação de novas análises.

Para além deste tipo de modelo, o valor também poderia ser representado através de uma rede de valor, definindo as relações entre indivíduos ou organizações que gerem valor tangível ou intangível. Um exemplo deste tipo de redes é a *Value Network Analysis*, permitindo especificar as interações existentes que criam ou não valor financeiro, internas ou externas. Com este modelo torna-se mais fácil explicar como criar valor a partir dos ativos existentes.

Outra forma de modelar o valor pode ser considerada utilizando modelos conceptuais de modelação de valor, nomeadamente o modelo conceptual de decomposição de valor para o cliente. É composto por diferentes fases, começando por dividir o valor em componentes mais simples (identificar tangíveis/intangíveis, ativos existentes/utilizados) a integrar no valor percebido. De seguida, recolhe-se informação da empresa relativa a um determinado período de tempo para que seja possível retirar conclusões sobre como o cliente compreende a proposição de valor apresentada. Por último, avalia-se a proposição de valor com os ativos de apoio. Para esta avaliação pode ser utilizada a Teoria dos Jogos, o método Multicritério AHP e o método *Fuzzy*.

Modelo de Canvas

<table>
<tr><th>Parceiros Chave</th><th>Atividades Chave</th><th>Proposta de Valor</th><th>Relacionamento com o Cliente</th><th>Segmentos de Cliente</th></tr>
<tr>
<td rowspan="3">Técnicos do departamento, no que toca a manutenção do hardware necessário (servidores)</td>
<td>Manutenção dos Servidores
Assistência técnica
Suporte ao que foi desenvolvido

Manutenção do data warehouse (carregamento de nova informação, atualização de dados)

Geração de reports</td>
<td rowspan="3">Armazém de dados com informação carregada centralizada

Informação segmentada
Aumento de Performance
Solução escalável
Probabilidade de erro de cálculos associados a avaliações dos alunos reduzidos

Acessibilidade na criação de análises baseadas no armazém de dados

Redução de custos na medida em que o utilizador consegue poupar tempo na criação e estudo dos resultados inerentes às análises que pretende</td>
<td>Assistência na realização de novos reports e novas funcionalidades</td>
<td rowspan="3">Clientes diretos/finais - Docentes</td>
</tr>
<tr><th>Recursos Chave</th><th>Canais</th></tr>
<tr>
<td>Alunos: responsáveis pelo desenvolvimento de novos reports/funcionalidades

Know how: propriedade intelectual criada

Servidores: permitir centralizar o sistema, permitir o trabalho por mais do que um aluno em paralelo.</td>
<td>Divulgação do armazem de dados criado aos docentes responsáveis pelo projeto

Divulgação do tipo de sistema criado a outros possíveis departamentos do ISEP interessados.</td>
</tr>
<tr><th colspan="2">Estruturas de Custo</th><th colspan="3">Fontes de Receita</th></tr>
<tr>
<td colspan="2">Manutenção do hardware (servidores e terminais onde o sistema vai estar disponível para utilização por parte dos docentes)</td>
<td colspan="3">Disponibilidade para criação de novos reports e funcionalidades

Disponibilidade para resolução de eventuais problemas, englobando resolução de problemas no data warehouse, no processo de carregamento de dados, nos possíveis reports gerados e a nível de manutenção de servidores.</td>
</tr>
</table>

Anexo B

Ambiente de desenvolvimento

A nível de equipamentos e ferramentas necessárias para o desenvolvimento da solução, foi disponibilizado um servidor do Departamento de Engenharia Informática do ISEP, devidamente preparado com o objetivo de permitir desenvolvimentos em paralelo. Neste ambiente, foi solicitado que constassem as seguintes ferramentas:

- ***SQL Management Studio*:** utilizado para a gestão das bases de dados e do AD;
- **MSSDT (*Microsoft SQL Server Data Tools*):** projetos criados do tipo SSIS (*SQL Server Integration Services*) que englobam a criação das bases de dados, dos processos de consulta às tabelas de configurações e dos processos de carregamento e limpeza de dados das fontes associada a cada uma das tabelas de *staging*, das dimensões e da tabela de factos. Também foi criado o projeto de SSAS, que permite a criação do cubo a partir do AD;
- **TFS (Team Foundation Server):** repositório de dados centralizado, utilizado para controlo de versões;
- **Power BI Desktop:** utilizado numa fase final do projeto para demonstrar a possibilidade do uso da solução integrada com uma aplicação de análise e apresentação de dados.

Numa primeira fase, foi construído um projeto por base de dados, com o objetivo de publicar automaticamente toda a estrutura de tabelas com base em *scripts* definidos. Na figura apresentada na secção Projeto MSSDT de publicação das bases de dados é brevemente apresentada uma estrutura parcial sendo, da esquerda para a direita, o projeto relativo à área de configurações, *staging*, e por último do AD.

No geral apresentam três áreas importantes, compostas por *scripts* para criação das tabelas relacionadas (pasta "Tables"), ficheiros de inserção de registos por defeito que é executado após toda a publicação da estrutura (definidos na pasta "PostDeploy", juntamente com o ficheiro de configuração[3]) e por último o conjunto de procedimentos que auxiliam todo o processo de extração e carregamento de dados criado (pasta "StoreProcedures"). À exceção dos procedimentos, os ficheiros possuem o nome da tabela juntamente com o nome da operação que é feita (criação ou inserção).

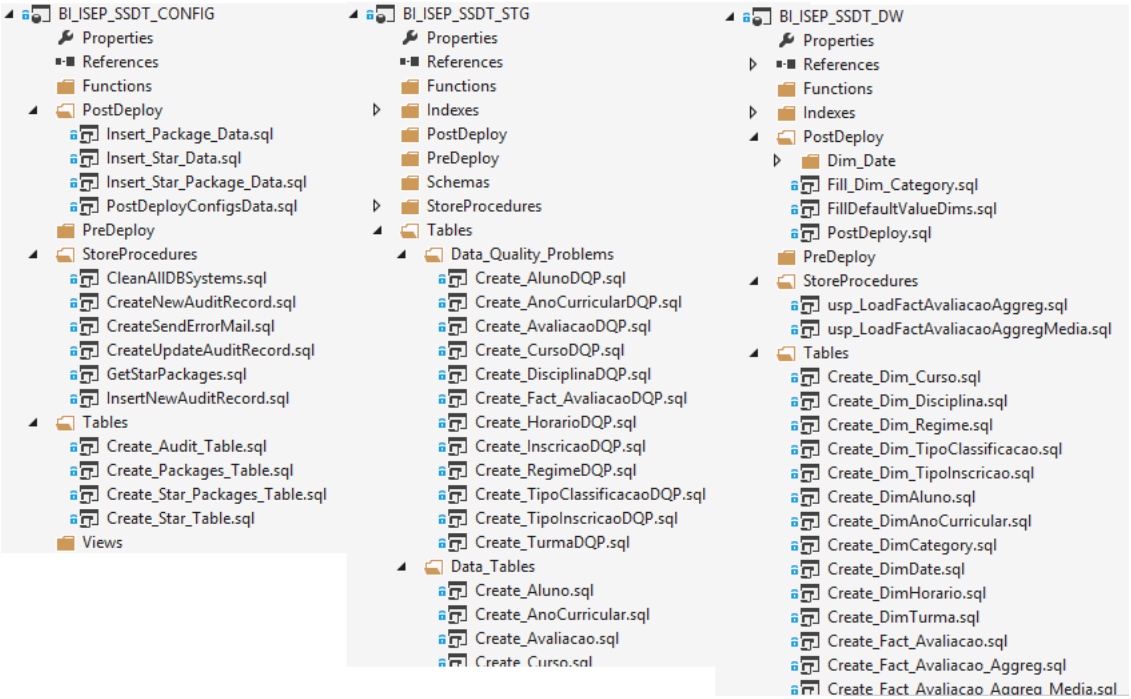
[3] Este ficheiros permite definir a ordem de execução dos ficheiros existentes na pasta de pós-publicação (*Post Deploy*).

De uma forma breve e relativamente aos procedimentos da área de configurações, é importante salientar que o procedimento denominado por “CleanAllDBSystems” foi criado exclusivamente para facilitar os testes e validação de dados no decorrer do desenvolvimento. Permite eliminar todos os registos de todas as bases de dados, mantendo apenas os valores por defeito necessários (como se fosse feita uma nova publicação). À exceção deste, todos os outros são utilizados pelos processos de extração e de carregamento, permitindo inserir/atualizar registos de auditoria de execução e ainda o envio de correio eletrónico para destinatários definidos, com os resultados da execução apresentados em forma de tabela.

Na estrutura de projeto seguinte, a estrutura de criação de tabelas encontra-se dividida entre tabelas de *staging* e tabelas de *staging* do tipo DQP. Por último, no projeto de criação do AD é de salientar o ficheiro de inserção de valores por defeito nas dimensões. Estes valores servem para serem mapeados quando não existe mapeamento com os dados do domínio. O *script* referido é relativo ao código apresentado na secção.

O passo seguinte passou por criar o projeto SSIS que engloba cada um dos pacotes de execução de extração e carregamento de dados para as tabelas/dimensões criadas, contendo a lógica de negócio específica de cada domínio (secção Estrutura geral do projeto SSIS e SSAS). Este processo começa a partir de um pacote principal, que consulta as configurações existentes para determinar quais os pacotes e por que ordem é que vão ser executados. Por fim, foi criado o projeto SSAS para criação do cubo de dados, sendo considerado apenas neste âmbito, o cubo denominado por “DEV DW”.

Projeto MSSDT de publicação das bases de dados



Conteúdo do ficheiro de inserção de registos por defeito

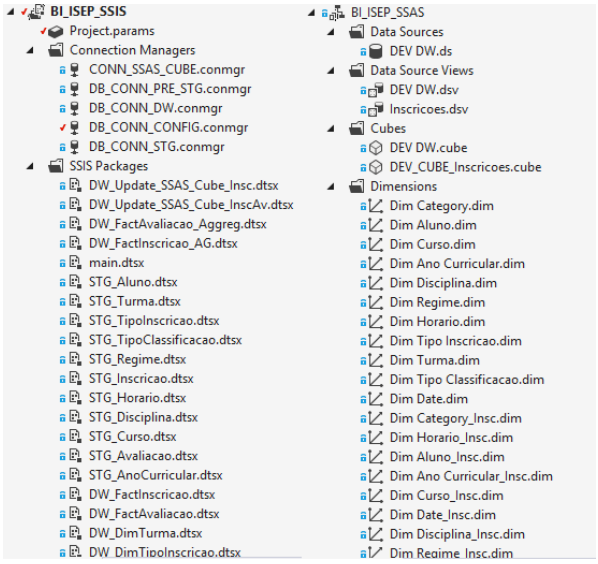
```
INSERT INTO [DEV_DW].[dbo].Dim_Curso([Sigla],[Descritivo],[EffectiveDate])
VALUES('-2', 'Not Applicable', GETDATE())

INSERT INTO
[DEV_DW].[dbo].Dim_Disciplina([Sigla],[Descritivo],[Ano],[Semestre],[ECTS],
[EffectiveDate],[IsCurrent])
VALUES ('-2','Not Applicable',0,0,0,GETDATE(),0)

INSERT INTO [DEV_DW].[dbo].Dim_Regime([Sigla],[Descritivo],[EffectiveDate])
VALUES ('-2','Not Applicable', GETDATE())

INSERT INTO
[DEV_DW].[dbo].Dim_TipoClassificacao([Sigla],[Descritivo],[EffectiveDate])
VALUES('-2', 'Not Applicable', GETDATE())
```

Estrutura geral do projeto SSIS e SSAS



Anexo C

Análises realizadas do ano letivo 2013-2014

Neste anexo são apresentadas as restantes análises criadas e respetiva comparação com as análises originais.

Contagem de classificações por semestre e disciplina

Este tipo de análise permite contabilizar o número de alunos que obteve determinado tipo de classificação como nota final, por disciplina e por semestre em que cada disciplina é lecionada. Os dados são provenientes da tabela de factos de avaliação dado que possui a granularidade pretendida (disciplina, semestre da disciplina e tipos de classificação).

Na análise original é possível verificar que se contabilizou um elevado número de registos total. É de notar que é correto que este total não seja igual ao total que foi sendo obtido nas análises anteriores (mil cento e oitenta e um registos) dado que a granularidade da pesquisa aumentou, o que reflete que um aluno obtém diferentes tipos de classificações às diferentes unidades curriculares. Este pode ser contabilizado mais do que uma vez, mas em tipos de classificações diferentes e disciplinas diferentes.

De uma forma breve, é possível verificar que para qualquer semestre/disciplina, a contabilização de alunos mais elevada foca-se nas aprovações. O número de classificações mais elevado seguinte vai variando entre reprovações, NF, SMNF, SMR e NC. A mais baixa é sem dúvida SMS. Relativamente ao semestre existe mais informação sobre classificações obtidas relativas às disciplinas do segundo e primeiro semestre. O valor mais baixo de contabilizações localiza-se no último semestre do curso.

No primeiro semestre, apesar da disciplina com mais classificações ser AMATA, é possível verificar que a unidade LAPR1 obteve mais aprovados em relação ao total de classificações, sendo que mais de metade das trezentas e vinte e nove classificações totais foram aprovações (apenas sessenta e quatro contabilizações para outros tipos de classificações). A unidade AMATA verifica-se com um maior número de classificações negativas.

Relativamente ao segundo, MATCP é a disciplina que obtém um maior número de classificações e faz parte das disciplinas com um pouco mais de metade de aprovações relativas ao total, juntamente com as unidades MDISC e PPROG. A unidade curricular ESOFT é o cenário mais desfavorável, com duzentas e cinquenta e uma classificações diferentes de aprovado.

No terceiro semestre, BDDAD é a unidade curricular com mais registos de aprovações, comparativamente ao total de classificações. Por outro lado, apesar de FISIAP ser a unidade

curricular que obtém um maior número de classificações, verifica-se a existência de um maior número de classificações negativas, sendo que foram encontradas duzentas e cinquenta e duas classificações diferentes de aprovado e apenas cento e oitenta e nove classificações de aprovação.

No que toca ao quarto semestre, a disciplina de LPROG é a que possui mais classificações obtidas. No entanto, as que se destacam são EAPLI, LAPR4 e SCOMP com um número de aprovações maior. Neste semestre, todas as unidades mantêm-se no mesmo cenário, sendo o número de classificações aprovadas superior às restantes.

Por fim, é possível verificar que os dois últimos semestres do curso possuem o mesmo tipo de cenário descrito no semestre anterior. No quinto semestre, SGRAI obtém um maior número de classificações aprovadas, apesar da unidade que tem mais registos de classificações ser ALGAV. Esta última é a que representa um maior número de classificações diferentes de aprovado. Relativamente ao último semestre, a unidade curricular IARTI tem um maior número de classificações e o maior número de aprovações. A disciplina de PESTI é a segunda unidade que possui mais registos, no entanto é a que mais contabilizou classificações diferentes de aprovado (noventa e nove registos).

Apenas se verificaram algumas alterações a nível de valores. No entanto, para melhor compreender detalhadamente as diferenças para a análise original, a comparação apresentada representa as diferenças de registos apenas para o que foi contabilizado na análise original.

Análise Original

Semestre	Disciplina	Aprovadc	Reprovado	SMR	SMS	SMNF	NC	NF	Grand Total
1	ALGAN	255	79				21	14	369
	AMATA	211	110				71	78	470
	APROG	216	23			71	4	37	351
	LAPR1	265	6			20		38	329
	PRCMP	228	12	28		73			341
1 Total		1175	230	28		164	96	167	1860
2	ESOFT	168	22			126		103	419
	LAPR2	230	80				13	92	415
	MATCP	269	141				19	37	466
	MDISC	246	100				27	69	442
	PPROG	288	13	12		53	6	70	442
2 Total		1201	356	12		179	65	371	2184
3	ARQCP	139	4	75		59		45	322
	BDDAD	205	14	69		13	30	30	361
	ESINF	148	5	113		24	18	35	343
	FSIAP	189	88	31		48	31	54	441
	LAPR3	213				28		12	253
3 Total		894	111	288		172	79	176	1720
4	EAPLI	215		32		15		45	307
	LAPR4	252				45			297
	LPROG	192	48	15		23	16	43	337
	RCOMP	181	10	77					268
	SCOMP	174	26	5			30		235
4 Total		1014	84	129		83	46	88	1444
5	ALGAV	174	14	24	50		4		266
	ARQSI	140	22	4	50		12		228
	ASIST	172	7	24					203
	GESTA	140	4				9		153
	LAPR5	141				16		23	180
	SGRAI	176	2	26	1	48			253
5 Total		943	49	78	101	64	25	23	1283
6	COMPA	153	51				31		235
	CORGA	149						15	164
	IARTI	177	42				46		265
	PESTI	141					99		240
6 Total		620	93				176	15	904
Grand Total		5847	923	535	101	662	487	840	9395

Análise Criada

Contagem de classificações por disciplina	Tipo de Classificação							
Semestre\Disciplina	APROV	NC	NF	REPRV	SMNF	SMR	SMS	Total Geral
1	**1171**	**98**	**167**	**233**	**164**	**27**		**1860**
ALGAN	253	22	14	80				369
AMATA	209	72	78	111				470
APROG	215	4	37	24	71			351
LAPR1	265		38	6	20			329
PRCMP	229			12	73	27		341
2	**1198**	**65**	**371**	**360**	**179**	**11**		**2184**
ESOFT	168		103	22	126			419
LAPR2	230	13	92	80				415
MATCP	267	19	37	143				466
MDISC	245	27	69	101				442
PPROG	288	6	70	14	53	11		442
3	**870**	**81**	**176**	**136**	**172**	**285**		**1720**
ARQCP	135		45	11	59	72		322
BDDAD	199	30	30	19	13	70		361
ESINF	145	18	35	9	24	112		343
FSIAP	178	33	54	97	48	31		441
LAPR3	213		12		28			253
4	**1000**	**46**	**88**	**102**	**84**	**125**		**1445**
EAPLI	213		45	2	15	32		307
LAPR4	252				45			297
LPROG	187	16	43	54	24	14		338
RCOMP	178			15		75		268
SCOMP	170	30		31		4		235
5	**921**	**26**	**23**	**64**	**64**	**82**	**103**	**1283**
ALGAV	162	5		21		27	51	266
ARQSI	138	12		23		4	51	228
ASIST	167			12		24		203
GESTA	140	9		4				153
LAPR5	141		23		16			180
SGRAI	173			4	48	27	1	253
6	**592**	**177**	**15**	**120**				**904**
COMPA	140	32		63				235
CORGA	149		15					164
IARTI	161	47		57				265
PESTI	142	98						240
Total Geral	**5752**	**493**	**840**	**1015**	**663**	**530**	**103**	**9396**

Comparação

Semestre	Disciplina	Aprovado	Reprovado	SMR	SMS	SMNF	NC	NF	Grand Total
1	ALGAN	-2	+1				+1		
	AMATA	-2	+1				+1		
	APROG	-1	+1						
	LAPR1								
	PRCMP	+1		-1					
1 Total		-4	+3	-1			+2		
2	ESOFT								
	LAPR2								
	MATCP	-2	+2						
	MDISC	-1	+1						
	PPROG		+1	-1					
2 Total		-3	+4	-1					
3	ARQCP	-4	+7	-3					
	BDDAD	-6	+5	+1					
	ESINF	-3	+4	-1					
	FSIAP	-11	+9				+2		
	LAPR3								
3 Total		-24	+25	-3			+2		
4	EAPLI	-2	+2						
	LAPR4								
	LPROG	-5	+6	-1		+1			+1
	RCOMP	-3	+5	-2					
	SCOMP	-4	+5	-1					
4 Total		-14	+18	-4		+1			+1
5	ALGAV	-12	+7	+3	+1		+1		
	ARQSI	-2	+1		+1				
	ASIST	-5	+5						
	GESTA								
	LAPR5								
	SGRAI	-3	+2	+1					
5 Total		-22	+15	+4	+2		+1		
6	COMPA	-13	+12				+1		
	CORGA								
	IARTI	-16	+15				+1		
	PESTI	+1					-1		
6 Total		-28	+27				+1		
Grand Total		-95	+92	-5	+2	+1	+6		+1

Contagem de classificações com nota de freq. positiva por semestre e disciplina

A análise original possui o mesmo âmbito que a análise anterior. Contabiliza o número de classificações obtidas apenas nos casos em que a nota de frequência obtida seja igual ou superior a dez valores. Da mesma forma, a tabela de factos de avaliação foi utilizada como fonte para esta análise, sendo apenas acrescentada na DSV do cubo e na tabela de factos de avaliação uma medida para contabilizar o registo com o valor um no caso de a frequência ser positiva. Caso contrário, contabiliza com zero. Assim é possível ser feito um somatório para compreender a quantidade de registos.

É possível verificar que o segundo e quarto semestres são os que contabilizam mais registos. Os tipos de classificação associado a aprovações possui o maior número de registos em todos os semestres do curso, ao contrário dos restantes tipos que vão variando de semestre para semestre.

Relativamente ao primeiro semestre, nas disciplinas de ALGAN e AMATA, todos os alunos contabilizados foram aprovados. A disciplina LAPR1 mantêm-se com um número de registos mais elevado e a que possui muito poucos registos com classificação negativa, sendo apenas um único. A unidade em falta, PRCMP, é a que possui um maior número de classificações negativas. Todas as unidades têm o valor de aprovações superior.

No segundo semestre, as unidades curriculares LAPR2 e MATCP apenas contabilizaram aprovações e a unidade curricular com mais registos é PPROG. O caso menos favorável reflete-se em ESOFT com um maior número de classificações diferentes de aprovado. No entanto, todas as unidades contabilizam um número de aprovações maior.

Relativamente ao terceiro semestre, LAPR3 possui um maior número de contabilizações e quase a totalidade são aprovações. Apenas dois registos com classificação SMNF. Neste caso a disciplina de ESINF contabiliza um número de aprovações baixo, sendo o valor de outros tipos de classificações mais elevado.

No que toca ao quarto semestre, é possível verificar que todos os registos contabilizam mais aprovações. A unidade curricular LAPR4 possui um maior número de registos e apenas contabilizou aprovações. Por outro lado, o pior cenário é LPROG com trinta e cinco registos com classificações diferentes de aprovado.

No quinto semestre, ASSIST é a unidade que tem o maior número de registos. Em todos os registos é possível verificar que existem mais aprovações do que outros tipos de classificações, destacando-se apenas LAPR5 e SGRAI como só tendo contabilizado aprovações.

Por fim, no último semestre apenas foram contabilizadas aprovações às três disciplinas consideradas, sendo que a que possui aprovações mais elevadas é CORGA e no outro extremo é IARTI. De notar que a unidade de PESTI não entrou para a contabilização pelo facto de não ser considerada nota de frequência.

Relativamente à análise apresentada e replicada com os dados do AD, é possível destacar logo que o número de registos total é mais elevado, sendo uma diferença de quatrocentos e trinta e cinco registos. Grande parte estão alocados ao número de aprovações e ao número de reprovações. Este número elevado apenas pode ser justificado com a fonte de dados que foi disponibilizada e utilizada para carregamento, sendo impossível descriminar de que forma é que a análise disponibilizada foi feita. Excetuando estes valores elevados, as restantes contagens possuem diferenças menos significativas.

No geral, esta análise possui o mesmo comportamento identificado e descrito sobre a análise original. O segundo e quarto semestre são igualmente os semestres com mais registos e o número de aprovações regista-se como sendo também o tipo de classificação mais obtido. Verifica-se também que a ordem das restantes classificações por número de registos contabilizados mantêm-se.

Foi possível concluir que o primeiro, segundo, quarto e sexto semestre mantêm-se às conclusões retiradas relativamente à análise original. Relativamente ao terceiro semestre, onde a unidade curricular LAPR2 deixou de contabilizar o maior número de registos, passando a ser BDDAD com duzentos e trinta e três alunos, onde mais de metade obteve aprovação. Todas as unidades curriculares passaram a ter maior número de aprovações do que outros tipos de classificação. O quinto semestre também apresenta diferenças, sendo que ASSIST deixou de ser

a unidade com mais registos, passando a ser ARQSI. Neste contexto, SGRAI passou a contabilizar um número de classificações diferente de aprovado, nomeadamente dezasseis registos.

Análise Original

Count of Número		AprovDisc						
Semestre	Disciplina	Aprovado	Reprovado	SMR	SMS	SMNF	NC	Grand Total
1	ALGAN	219						219
	AMATA	77						77
	APROG	186	5			1	1	193
	LAPR1	245	1					246
	PRCMP	202		19				221
1 Total		929	6	19		1	1	956
2	ESOFT	137	3			14		154
	LAPR2	181						181
	MATCP	241						241
	MDISC	141	3					144
	PPROG	261	3	6			4	274
2 Total		961	9	6		14	4	994
3	ARQCP	90		46				136
	BDDAD	107		43			15	165
	ESINF	69	1	73			6	149
	FSIAP	141	8	24			1	174
	LAPR3	210				2		212
3 Total		617	9	186		2	22	836
4	EAPLI	162		31				193
	LAPR4	252						252
	LPROG	125	20	13			2	160
	RCOMP	151	1	27				179
	SCOMP	124	6	5			1	136
4 Total		814	27	76			3	920
5	ALGAV	112	1	12			4	129
	ARQSI	82	11	4			11	108
	ASIST	134	2	5				141
	GESTA	98					1	99
	LAPR5	137						137
	SGRAI	121		16				137
5 Total		684	14	37			16	751
6	COMPA	68						68
	CORGA	143						143
	IARTI	64						64
6 Total		275						275
Grand Total		4280	65	324	0	17	46	4732

Análise Criada

Contagem de classificações com nota positiva por tipo de classificação	Tipo de Classificação						
Semestre\Disciplina	APROV	NC	REPRV	SMNF	SMR	SMS	Total Geral
1	**904**	**1**	**7**	**1**	**16**		**929**
ALGAN	195	0	0				195
AMATA	65	0	0				65
APROG	194	1	5	1			201
LAPR1	237		1	0			238
PRCMP	213		1	0	16		230
2	**936**	**4**	**7**	**10**	**4**		**961**
ESOFT	149		2	10			161
LAPR2	176	0	0				176
MATCP	230	0	0				230
MDISC	125	0	1				126
PPROG	256	4	4	0	4		268
3	**780**	**22**	**45**	**0**	**177**		**1024**
ARQCP	127		6	0	38		171
BDDAD	170	15	5	0	43		233
ESINF	135	6	5	0	73		219
FSIAP	149	1	29	0	23		202
LAPR3	199			0			199
4	**964**	**2**	**48**	**0**	**69**		**1083**
EAPLI	210		2	0	31		243
LAPR4	252			0			252
LPROG	170	1	31	0	13		215
RCOMP	171		1		21		193
SCOMP	161	1	14		4		180
5	**829**	**15**	**26**	**0**	**37**	**0**	**907**
ALGAV	145	3	3		12	0	163
ARQSI	132	11	22		4	0	169
ASIST	154		1		5		160
GESTA	116	1	0				117
LAPR5	138			0			138
SGRAI	144		0	0	16	0	160
6	**263**	**0**	**0**				**263**
COMPA	68	0	0				68
CORGA	148						148
IARTI	47	0	0				47
PESTI	0	0					0
7	**0**						**0**
P/EST-ERASMUS 20	0						0
P/EST-ERASMUS 30	0						0
Total Geral	**4676**	**44**	**133**	**11**	**303**	**0**	**5167**

Comparação

Semestre	Disciplina	Aprovado	Reprovad	SMR	SMS	SMNF	NC	Grand Total
1	ALGAN	-24						-24
	AMATA	-12						-12
	APROG	+8						+8
	LAPR1	-8						-8
	PRCMP	+11	+1	-3				+9
1 Total		-25	+1	-3				-27
2	ESOFT	+12	-1			-4		+7
	LAPR2	-5						-5
	MATCP	-11						-11
	MDISC	-16	-2					-18
	PPROG	-5	+1	-2				-6
2 Total		-25	-2	-2		-4		-33
3	ARQCP	+37	+6	-8				+35
	BDDAD	+63	+5					+68
	ESINF	+66	+4					+70
	FSIAP	+8	+21	-1				+28
	LAPR3	-11				-2		-13
3 Total		+163	+36	-9		-2		+188
4	EAPLI	+48	+2					+50
	LAPR4							
	LPROG	+45	+11				-1	+55
	RCOMP	+20		-6				+14
	SCOMP	+37	+8	-1				+44
4 Total		+150	+21	-7			-1	+163
5	ALGAV	+33	+2				-1	+34
	ARQSI	+50	+11					+61
	ASIST	+20	-1					+19
	GESTA	+18						+18
	LAPR5	+1						+1
	SGRAI	+23						+23
5 Total		+145	+12				-1	+156
6	COMPA							
	CORGA	+5						+5
	IARTI	-17						-17
6 Total		-12						-12
Grand Total		+396	+68	-21		-6	-2	+435

Aprovações por ano curricular

Esta análise contabiliza o número de alunos que obteve um determinado número de aprovações por ano curricular. Foi disponibilizada uma análise relativa a totais de aprovações por ano letivo, ao qual vai ser utilizada para termo de comparação da análise corrente.

Verifica-se que a nível de totais por aprovações, mantêm-se os mesmos valores apresentados na análise anterior (inscrições versus aprovações). No entanto, a nível de somatórios por ano curricular, verifica-se a existência de informação de mais alunos do primeiro ano, do terceiro e por último do segundo ano curricular.

Em específico para inscrições no primeiro ano curricular, verifica-se um maior número em zero e as dez unidades curriculares aprovadas com setenta alunos. Seguem-se como mais elevadas as nove e as duas/três unidades curriculares aprovadas. O valor mais baixo obtido foi nas cinco unidades curriculares com vinte alunos.

Relativamente ao segundo ano curricular, o valor mais alto situa-se nas dez unidades aprovadas, identificando cinquenta e três alunos. Seguem-se quatro/cinco unidades aprovadas, com quarenta e sete alunos. Apenas dois alunos obtiveram aprovação a treze unidades curriculares. No entanto e por último, no terceiro ano curricular, quarenta e nove alunos obtiveram

aprovação a uma única unidade. A segunda maior contagem verifica-se para as dez unidades curriculares aprovadas com mais quarenta e sete alunos. Por outro lado, registam-se valores baixos na aprovação de doze/treze unidades.

Com o objetivo de conseguir comparar com as análises fornecidas, a fonte de dados utilizada para esta análise foi a tabela de agregação de avaliação, sendo as linhas compostas pela informação dos anos e as colunas com informação do número de aprovações. Nesta análise, o número total de alunos possui uma diferença mínima de seis registos a mais quando comparado com a análise original. Este número de registos deve-se à granularidade da tabela de factos agregada utilizada, dado que pode originar mais do que um registo por aluno. Como exemplo teríamos o caso de uma avaliação obtida no âmbito de um determinado regime e outra avaliação diferente onde não existe informação sobre qual foi o regime. Teríamos uma contagem de aprovados por dois tipos de regime distintos.

É possível verificar que as contagens não saem do espectável, sendo as diferenças existentes menores. O número de unidades curriculares aprovadas mais e menos afluente mantêm-se igual, mas com uma diferença menor a nível de contagens.

Análise Original

	Número de UCs em que o aluno obteve aprovação															
Ano do Alun	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1	70	34	35	26	25	20	28	29	30	48	70					415
2	40	23	28	28	47	47	29	31	29	17	53	3		2		377
3	59	49	43	39	25	30	24	20	13	15	47	19	4	2		389
Total	169	106	106	93	97	97	81	80	72	80	170	22	4	4	0	1181

Análise Criada

Contagem de Alunos	Número de Aprovações														
Ano Curricular\Ano Entrada	0	1	2	3	4	5	6	7	8	9	10	11	13	12	Total Geral
1	74	34	35	26	26	19	26	31	30	48	69				418
2	40	24	30	26	50	48	28	30	29	17	51	3	1		377
3	66	51	36	47	24	31	24	19	14	17	47	10	2	4	392
Total Geral	180	109	101	99	100	98	78	80	73	82	167	13	3	4	1187

Comparação

Ano do Aluno	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
1	+4				+1	-1	-2	+2			-1				+3
2		+1	+2	-2	+3	+1	-1	-1			-2		+1	-2	
3	+7	+2	-7	+8	-1	+1		-1	+1	+2		-9	-2	+2	+3
Total	+11	+3	-5	+6	+3	+1	-3		+1	+2	-3	-9	-1		+6

Reprovações por ano curricular

A análise apresentada possui o mesmo âmbito da anterior, alterando-se apenas o domínio para reprovações por ano curricular. Contabiliza todas as classificações finais obtidas, cujo tipo de classificação seja apenas reprovado.

A nível de totais por ano curricular, é possível observar que os valores se mantiveram nas análises originais, sendo o primeiro ano o que é constituído por mais contagens, seguindo-se do terceiro e por último o segundo ano curricular.

Já no que toca ao número de unidades curriculares reprovadas verifica-se que metade dos alunos não obteve qualquer reprovação. A contabilização relativa ao primeiro ano curricular traduz que nenhum dos alunos obteve reprovações. Nos restantes anos, este número de unidades curriculares mantêm-se mais alto, descendo de valor à medida que o número de unidades curriculares reprovadas cresce.

Relativamente à análise comparativa criada, a fonte de dados utilizada é a mesma tabela de avaliação agregada e utiliza o valor dos anos para as linhas e o número de unidades para as colunas da tabela.

Dado que foi utilizada a mesma tabela de factos agregada que a análise anterior, a diferença de total é exatamente a mesma. Relativamente ao primeiro ano, verificou-se que afinal foram contabilizados mais alunos com reprovações a unidades, refletindo-se apenas numa diferença de três registos de alunos no final. No caso do segundo, foram contabilizados mais alunos como não tendo reprovações a nenhuma unidade e/ou a uma, sendo que as restantes contagens reduziram. Por último, no terceiro ano verificou-se um aumento de alunos com reprovações a nenhuma unidade, sendo que as restantes contagens sofreram pela negativa.

Análise Original

	Nº de UCs em atraso							
Ano	0	1	2	3	4	5	6	Total
1	415							415
2	137	92	52	54	29	10	3	377
3	133	112	72	36	30	6		389
Total	685	204	124	90	59	16	3	1181

Análise Criada

Contagem de Alunos	Número de Reprovações							
Ano Curricular\Ano Entrada	0	1	2	3	4	5	6	Total Geral
1	200	87	56	42	25	5	3	418
2	190	112	48	19	5	2	1	377
3	237	94	41	15	2	2	1	392
Total Geral	627	293	145	76	32	9	5	1187

Comparação

Ano Aluno	0	1	2	3	4	5	6	Total
1	-215	+87	+56	+42	+25	+5	+3	+3
2	+53	+20	-4	-35	-24	-8	-2	
3	+104	-18	-31	-21	-28	-4	+1	+3
Total	-58	+89	+21	-14	-27	-7	+2	+6

Aprovações por ano curricular e regime

A análise seguinte permite verificar o número de alunos que obtiveram um determinado número de aprovações, por ano curricular e regime. Esta análise é idêntica à análise sobre aprovações anteriormente apresentada, incluindo mais um parâmetro granular onde é possível verificar a situação de alunos aprovados por regime ao qual frequentaram as unidades. A fonte de dados é igualmente provenientes da tabela de factos de avaliação agregada.

De uma forma geral, a nível de totais por aprovações e por ano curricular, os valores totais mantiveram-se. Os regimes integral e parcial são os que contabilizam o maior número de alunos. Nos restantes verifica-se a existência de muito poucos alunos. Relativamente ao primeiro ano, é possível compreender que mais de metade dos alunos frequentou as unidades aprovadas em regime integral, sendo que sessenta e nove obtiveram aprovação a dez unidades curriculares. O valor mais baixo verifica-se na aprovação de apenas cinco unidades curriculares. Por outro lado, o regime parcial é o segundo total mais alto, contabilizando apenas com trinta e três alunos. Os regimes extraordinário, extra-curricular e gratuito contabilizam muito poucos alunos, sendo o número máximo de aprovações obtido nas zero unidades curriculares. Apenas um aluno esteve inscrito em ERASMUS e obteve aprovação a sete unidades e foram encontrados três alunos que não tem informação sobre o regime frequentado (*Not Applicable*).

Relativamente ao segundo ano curricular, apenas foi encontrada informação sobre o regime integral, mobilidade e parcial. No integral, verifica-se igualmente uma frequência de mais de metade do total do ano, sendo que o número de unidades mais aprovadas não se alterou das dez unidades. No entanto, o valor mais baixo de unidades aprovadas alterou-se para as treze unidades curriculares com apenas um aluno. Os valores mais altos e baixos mantêm-se no regime parcial. Apenas foi encontrado um aluno com uma frequência no regime de imobilidade e ao qual ficou aprovado a quatro unidades.

Por fim, o último ano curricular mantêm-se nos mesmos níveis descritos para os anos curriculares anteriores. Por outro lado, o número de alunos em regime parcial é mais elevado que nos anos anteriores. Foram encontrados quatro registos sem informação de regime e um deles possui uma unidade aprovada. Foi encontrado também um aluno que frequentou o regime Vasco da Gama e ao qual não obteve aprovação.

Análise Criada

Contagem de Alunos	Número de Aprovações														
Ano Curricular\Regime	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Total Geral
1	**74**	**34**	**35**	**26**	**26**	**19**	**26**	**31**	**30**	**48**	**69**				**418**
Erasmus			1												1
Extra-Curricular	4	3	2	1		1		1							12
Extraordinário	7		1	1			1								10
Gratuito	2	1	1												4
Integral	44	23	25	22	23	16	25	30	30	48	69				355
Not Applicable	3														3
Parcial	14	7	5	2	3	2									33
2	**40**	**24**	**30**	**26**	**50**	**48**	**28**	**30**	**29**	**17**	**51**	**3**		**1**	**377**
Integral	21	16	23	20	37	34	21	30	29	17	51	3		1	303
Mobilidade					1										1
Parcial	19	8	7	6	12	14	7								73
3	**66**	**51**	**36**	**47**	**24**	**31**	**24**	**19**	**14**	**17**	**47**	**10**	**4**	**2**	**392**
Erasmus	1	2		2											5
Integral	16	9	12	17	14	18	19	17	14	17	47	10	4	2	216
Not Applicable	3	1													4
Parcial	45	39	24	28	10	13	5	2							166
Vasco da Gama	1														1
Total Geral	**180**	**109**	**101**	**99**	**100**	**98**	**78**	**80**	**73**	**82**	**167**	**13**	**4**	**3**	**1187**

Aprovações por ano curricular e horário

Esta análise verifica o número de alunos que obtiveram um determinado número de aprovações apenas por horário. Os dados são igualmente provenientes da tabela agregada de avaliação, o que significa que os totais se mantêm.

Relativamente aos dados do primeiro ano curricular, é possível verificar que o horário laboral é o escolhido por mais de metade da totalidade. As contagens mais altas verificam-se para as zero e as dez unidades curriculares aprovadas. Como segundo horário mais afluente é apresentado o pós-laboral, contendo o número de alunos mais elevado nas zero unidades aprovadas. Por outro lado, foram encontrados nove alunos cuja informação sobre o horário não é conhecida, sendo identificados na coluna denominada por “Não inscrito”.

No segundo ano curricular, o plano laboral mantêm-se no patamar mais elevado, sendo que o maior número de alunos situa-se entre as cinco e as dez unidades aprovadas. O pós-laboral manteve a tendência anterior. No entanto, apenas foi encontrada informação sobre um aluno ao qual o horário não é conhecido.

Por fim, o último ano mantém o mesmo comportamento e apresenta uma maior aproximação de alunos com horário laboral e pós-laboral. Foram também encontrados registos sem informação de horário, nomeadamente informação sobre três alunos.

Análise Criada

Contagem de Alunos	Número de Aprovações													
Ano Curricular\Horário	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Total Geral
1	**74**	**34**	**35**	**26**	**26**	**19**	**26**	**31**	**30**	**48**	**69**				**418**
Laboral	45	27	28	23	23	13	22	24	24	46	65				340
Não Inscrito	8					1									9
Pós-Laboral	21	7	7	3	3	5	4	7	6	2	4				69
2	**40**	**24**	**30**	**26**	**50**	**48**	**28**	**30**	**29**	**17**	**51**	**3**		**1**	**377**
Laboral	16	16	23	20	38	39	23	27	27	13	46	2		1	291
Não Inscrito					1										1
Pós-Laboral	24	8	7	6	11	9	5	3	2	4	5	1			85
3	**66**	**51**	**36**	**47**	**24**	**31**	**24**	**19**	**14**	**17**	**47**	**10**	**4**	**2**	**392**
Laboral	24	24	21	25	20	27	21	14	13	16	41	10	3	2	261
Não Inscrito	1	2													3
Pós-Laboral	41	25	15	22	4	4	3	5	1	1	6		1		128
Total Geral	**180**	**109**	**101**	**99**	**100**	**98**	**78**	**80**	**73**	**82**	**167**	**13**	**4**	**3**	**1187**

Anexo D

Análises Originais – Ano Letivo 2012-2013

Neste anexo são apresentadas as análises disponibilizadas referentes ao ano letivo 2012-2013.

Unidades curriculares inscritas versus unidades aprovadas

	Número de UCs em que o aluno obteve aprovação															
Nº UCs *	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1	19	37														56
2	16	5	14													35
3	16	3	5	6												30
4	15	5	5	2	7											34
5	28	5	9	7	7	13										69
6	20	10	13	7	14	10	16									90
7	18	8	11	6	13	11	9	8								84
8	16	9	5	6	10	11	13	12	8							90
9	12	7	8	6	12	11	11	12	12	23						114
10	25	16	14	15	14	20	21	27	28	45	142					367
11	3	2	1	3	1	9	4	13	12	15	8	21				92
12		2		3	1	2	7	6	6	2	8	7	6			50
13					1	2	2	2	2		3		1	1		14
14															1	1
15																0
Total	188	109	85	61	80	89	83	80	68	85	161	28	7	1	1	1126

Unidades curriculares inscritas versus unidades com reprovações

	Nº de UCs em atraso							
Nº UCs	0	1	2	3	4	5	6	Total
1	41	14						55
2	20	10	5					35
3	10	11	8	1				30
4	13	12	7	1	1			34
5	33	16	6	3	5	6		69
6	44	15	11	10	8	2		90
7	29	13	20	18	3	1		84
8	44	12	16	10	7	1		90
9	48	18	13	22	12	1		114
10	325	17	6	8	10		1	367
11	5	45	28	4	9	1		92
12	1	3	17	18	11			50
13				7	7			14
14					1			1
15								0
Total	613	186	137	102	74	12	1	1125

Aprovações por ano curricular

	Número de UCs em que o aluno obteve aprovação															
Ano do Aluno	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1	67	31	24	14	29	29	25	23	23	31	71					367
2	48	22	26	29	28	39	31	28	24	29	36	9	2			351
3	73	56	35	18	23	21	27	29	21	25	54	19	5	1	1	408
Total	188	109	85	61	80	89	83	80	68	85	161	28	7	1	1	1126

Reprovações por ano curricular

	Nº de UCs em atraso							
Ano	0	1	2	3	4	5	6	Total
1	367							367
2	107	86	63	52	34	8	1	351
3	139	100	74	50	40	4		407
Total	613	186	137	102	74	12	1	1125

Contagem de classificações por semestre e disciplina

Semestre	Disciplina	Aprovadc	Reprovado	SMR	SMS	SMNF	NC	NF	Grand Total
1	ALGAN	224	102				19	16	361
	AMATA	158	136				49	60	403
	APROG	215	37	1		49	5	24	331
	LAPR1	230	1			14		50	295
	PRCMP	221	5	16		69			311
1 Total		1048	281	17		132	73	150	1701
2	ESOFT	240	60				33	83	416
	LAPR2	234	24			55		72	385
	MATCP	222	105				51	53	431
	MDISC	174	25			90	5	73	367
	PPROG	225	24	1		97		79	426
2 Total		1095	238	1		242	89	360	2025
3	ARQCP	206	12	102				30	350
	BDDAD	135	8	93		14	30	27	307
	ESINF	230	5	55		11	16	2	319
	FSIAP	163	134	15			37	85	434
	LAPR3	165				16	1	10	192
3 Total		899	159	265		41	84	154	1602
4	EAPLI	158	4	18		32		35	247
	LAPR4	145	5			47		24	221
	LPROG	155	48	6		1	46	30	286
	RCOMP	187	6	66					259
	SCOMP	190	25	5			34		254
4 Total		835	88	95		80	80	89	1267
5	ALGAV	152	36	10			30		228
	ARQSI	195	22	5	48		7		277
	ASIST	170	18	13			29		230
	GESTA	207	6	8			15		236
	LAPR5	185				14			199
	SGRAI	159		26		35			220
5 Total		1068	82	62	48	49	81		1390
6	COMPA	170	47				47		264
	CORGA	182		1			1	27	211
	IARTI	155	68				43		266
	PESTI	164	2				93		259
6 Total		671	117	1			184	27	1000
Grand Total		5616	965	441	48	544	591	780	8985

Anexo E

Análises Criadas – Ano Letivo 2012-2013

Neste anexo são apresentadas as análises criadas referentes ao ano letivo 2012-2013.

Unidades curriculares inscritas versus unidades aprovadas

Contagem de Alunos	Número de Aprovações													
Número de Inscrições	0	1	2	3	4	5	6	7	8	9	10	11	12	14	Total Geral
1	30	32													62
2	26	15	18												59
3	26	11	11	8											56
4	26	10	19	10	17										82
5	27	13	11	31	17	34									133
6	17	12	15	15	18	20	20								117
7	12	15	8	7	11	15	8	19							95
8	10	9	4	11	11	8	15	17	11						96
9	7	4	4	7	4	4	5	11	12	11					69
10	6	15	10	9	10	16	20	21	20	41	103				271
11			1	2	1	3	3	7	6	8	4	4			39
12					1	2	5	2		2	3	1	2		18
13						1							1		2
14														1	1
Total Geral	187	136	101	100	90	103	76	77	49	62	110	5	3	1	1100

Unidades curriculares inscritas versus unidades com reprovações

Contagem de Alunos	Número de Reprovações							
Número de Inscrições	0	1	2	3	4	5	6	Total Geral
1	58	4						62
2	48	11						59
3	38	15	2	1				56
4	58	19	5					82
5	75	41	15	2				133
6	57	35	17	5	3			117
7	42	23	24	5		1		95
8	38	26	20	7	4	1		96
9	31	18	13	4	2	1		69
10	134	55	46	19	9	6	2	271
11	13	12	5	7	1	1		39
12	4	6	3	3	2			18
13		1	1					2
14	1							1
Total Geral	597	266	151	53	21	10	2	1100

Aprovações por ano curricular

Contagem de Alunos	Número de Aprovações														
Ano Curricular\Ano Entrada	0	1	2	3	4	5	6	7	8	9	10	11	12	14	Total Geral
1	66	44	23	28	24	30	20	27	22	26	56				366
2	54	39	42	50	37	34	27	23	15	12	15	2	1		351
3	72	53	37	22	30	39	28	27	12	24	39	3	2	1	389
Total Geral	**192**	**136**	**102**	**100**	**91**	**103**	**75**	**77**	**49**	**62**	**110**	**5**	**3**	**1**	**1106**

Reprovações por ano curricular

Contagem de Alunos	Número de Reprovações							
Ano Curricular\Ano Entrada	0	1	2	3	4	5	6	Total Geral
1	165	89	70	22	11	7	2	366
2	197	88	43	16	5	2		351
3	241	89	38	15	5	1		389
Total Geral	**603**	**266**	**151**	**53**	**21**	**10**	**2**	**1106**

Aprovações por ano curricular e regime

Contagem de Alunos	Número de Aprovações														
Ano Curricular\Regime	0	1	2	3	4	5	6	7	8	9	10	11	12	14	Total Geral
1	**66**	**44**	**23**	**28**	**24**	**30**	**20**	**27**	**22**	**26**	**56**				**366**
Extra-Curricular	2														2
Extraordinário		1			1										2
Integral	46	35	19	28	19	29	20	27	22	26	56				327
Parcial	18	8	4		4	1									35
2	**54**	**39**	**42**	**50**	**37**	**34**	**27**	**23**	**15**	**12**	**15**	**2**	**1**		**351**
Gratuito			1												1
Integral	29	30	28	41	33	34	25	23	15	12	15	2	1		288
Not Applicable	3		1												4
Parcial	22	9	12	9	4		2								58
3	**72**	**53**	**37**	**22**	**30**	**39**	**28**	**27**	**12**	**24**	**39**	**3**	**2**	**1**	**389**
Erasmus		1	1												2
Extra-Curricular	2	2													4
Integral	17	11	18	15	24	36	28	27	11	24	39	3	2	1	256
Not Applicable	2														2
Parcial	51	39	18	7	6	3			1						125
Total Geral	**192**	**136**	**102**	**100**	**91**	**103**	**75**	**77**	**49**	**62**	**110**	**5**	**3**	**1**	**1106**

Aprovações por ano curricular e horário

Contagem de Alunos	Número de Aprovações													
Ano Curricular\Horário	0	1	2	3	4	5	6	7	8	9	10	11	12	14	Total Geral
1	66	44	23	28	24	30	20	27	22	26	56				366
Laboral	46	34	16	23	15	25	17	25	20	25	53				299
Não Inscrito		1			1										2
Pós-Laboral	20	9	7	5	8	5	3	2	2	1	3				65
2	54	39	42	50	37	34	27	23	15	12	15	2	1		351
Laboral	35	23	29	37	31	28	22	21	12	11	13	2	1		265
Não Inscrito		1				1									2
Pós-Laboral	19	15	13	13	6	5	5	2	3	1	2				84
3	72	53	37	22	30	39	28	27	12	24	39	3	2	1	389
Laboral	26	16	19	18	23	28	20	25	9	21	33	3	2	1	244
Não Inscrito		1													1
Pós-Laboral	46	36	18	4	7	11	8	2	3	3	6				144
Total Geral	192	136	102	100	91	103	75	77	49	62	110	5	3	1	1106

Contagem de classificações por semestre e disciplina

Contagem de classificações por disciplina	Tipo de Classificação							
Semestre\Disciplina	APROV	NC	NF	REPRV	SMNF	SMR	SMS	Total Geral
1	1248	71	147	287	132	17		1902
ALGAN	419	18	13	105				555
AMATA	157	48	60	139				404
APROG	218	5	24	37	49	1		334
LAPR1	233		50	1	14			298
PRCMP	221			5	69	16		311
2	1086	88	359	248	242	1		2024
ESOFT	237	33	83	62				415
LAPR2	234		72	25	55			386
MATCP	219	50	53	108				430
MDISC	174	5	73	26	90			368
PPROG	222		78	27	97	1		425
3	874	88	154	177	41	271		1605
ARQCP	206		30	7		107		350
BDDAD	134	30	27	10	14	92		307
ESINF	232	16	2	7	11	53		321
FSIAP	137	41	85	153		19		435
LAPR3	165	1	10		16			192
4	833	80	89	92	80	93		1267
EAPLI	158		35	5	32	17		247
LAPR4	145		24	5	47			221
LPROG	154	46	30	49	1	6		286
RCOMP	187			6		65		258
SCOMP	189	34		27		5		255
5	1041	82		109	49	62	49	1392
ALGAV	134	30		54		10		228
ARQSI	190	7		26		5	49	277
ASIST	167	29		22		13		231
GESTA	205	16		7		8		236
LAPR5	186				14			200
SGRAI	159				35	26		220
6	642	186	27	140		1		996
COMPA	156	46		60				262
CORGA	182	1	27			1		211
IARTI	141	44		80				265
PESTI	163	95						258
Total Geral	5724	595	776	1053	544	445	49	9186

Contagem de exames realizados por ano letivo

Total Exames Época Normal	Total Exames Época Recurso	Total Exames Época Especial
5093	1985	96

Média de notas de alunos finalistas

Contagem de Alunos	Média de notas							
Número de Anos	11	12	13	14	15	16	17	Total Geral
2			1					1
3		5	18	7	5	4	3	42
4		6	11	6	1	1		25
5		9	7	2	1			19
6		13	4		1			18
7	1	4	1					6
8		1						1
Total Geral	**1**	**38**	**42**	**15**	**8**	**5**	**3**	**112**

Média de notas de alunos

Contagem de Alunos	Média de notas								
Número de Anos	10	11	12	13	14	15	16	17	18	Total Geral
1	4	6	12	6	4	2			1	35
2	1	7	9	3	1	1				22
3		8	18	31	12	7	5	3		84
4		4	16	19	7	1	1			48
5		3	18	9	4	1				35
6		2	18	6	2	1				29
7		2	5	2						9
8			1							1
Total Geral	**5**	**32**	**97**	**76**	**30**	**13**	**6**	**3**	**1**	**263**

Resumo dos Problemas de Qualidade de Dados

Geral - Ano Letivo 2012-2013	
Motivo	**Quantidade de registos**
Registos com a época inválida (“C2”,”DZ”,”PO”)	403
Registos duplicados	16
Registos com nota final negativa	26
Registos encontrados durante o processamento que, ao serem comparados com os dados das notas de aluno inseridos na tabela de *staging*, foram encontrados registos para o mesmo aluno, para a mesma disciplina e para o mesmo ano letivo. No entanto a nota final não corresponde à que já foi inserida.	319
Registos sem mapeamento na dimensão aluno	13