



Universidade de Lisboa  
Faculdade de Ciências  
Mestrado em Bioinformática e Biologia Computacional - MBBC  
Mestrado em Informática - MI

Integração e Processamento Analítico de Informação - IPAI

**PROJECTO:**  
***Audiências Televisivas***

Docente: António Ferreira

André Oliveira, 45648 – MI  
Maria Móteiro, 43178 - MBBC  
Tânia Maldonado, 44745 - MBBC

14 de Maio de 2017

## Índice

1.	Introdução .....	2
2.	Análise das Fontes de Dados .....	3
2.1.	Espetadores.....	3
2.1.1.	Tratamento de Dados .....	3
2.2.	Tipos de Programas Televisivos.....	7
2.3.	Canais Vistos pelos Espetadores.....	9
2.3.1.	Tratamento de Dados .....	9
2.4.	Programação dos Canais Televisivos.....	13
2.4.1.	Tratamento de Dados .....	14
2.5.	Classes Sociais dos Espetadores .....	16
2.6.	Fonte de Dados Adicionais .....	16
3.	Relação entre Fontes de Dados.....	17
4.	Modelação Dimensional.....	18
4.1.	Processo de Negócio .....	18
4.2.	Perguntas Analíticas .....	18
4.3.	Definição do Grão.....	18
4.4.	Dimensões do Negócio .....	19
4.4.1.	Dimensão Espetador .....	19
4.4.2.	Dimensão Programa.....	20
4.4.3.	Dimensão Data .....	21
4.4.4.	Dimensão Horário .....	21
4.5.	Registo de Mudanças Lentas .....	22
4.6.	Medidas Numéricas .....	23
4.7.	Diagrama em Estrela do <i>Data Warehouse</i> .....	24
5.	Sistema ETL.....	25
5.1.	Extração dos Dados .....	26
5.2.	Transformação dos Dados.....	27
5.3.	Carregamento do Data Warehouse .....	29
5.3.1.	SQL Server Management Studio.....	29
5.3.2.	SQL Server Business Intelligence Development Studio .....	31
5.3.2.1.	Integration Services .....	31
5.3.2.2.	Analysis Services .....	32
6.	Conclusão .....	35
7.	Bibliografia .....	36
8.	Anexos.....	37

## 1. Introdução

Os dados das audiências são um importante elemento de gestão das estações televisivas, estando na base da definição do preço da publicidade a cobrar aos anunciantes, assim como da tomada de decisões sobre as grelhas de programação, entre outros.

O projeto desta unidade curricular envolve a modelação e construção de um *data warehouse* que incorpore dados de audiências televisivas no período de tempo que vai desde o dia 1 de janeiro de 1996, até 30 de junho de 1996. Este *data warehouse* será, idealmente, capaz de dar resposta a um leque de cenários de tomada de decisão, mas neste contexto será relativo a um único processo de negócio.

Propusemo-nos, no presente relatório, traçar algumas tendências e hábitos de consumo de televisão durante o primeiro semestre de 1996. Na primeira fase do projeto foi feita uma análise dos dados existentes e do negócio, com o objetivo de compreender as fontes de dados disponibilizadas assim como as interligações entre as diversas fontes de dados e, por fim, descrever um processo de negócio que possa utilizar estes dados. O processo de negócio escolhido pretende analisar os programas visto por um espetador durante um período de tempo.

Na segunda fase do trabalho, procedemos à modelação dimensional do *data warehouse*, onde foram definidos o grão e o tipo de tabela de factos, bem como as dimensões e medidas envolvidas, por forma a preparar a construção do sistema ETL.

O dito sistema ETL foi construído e demonstrado na terceira fase do projeto, permitindo já compor um relatório analítico a título de exemplo e preparação para uma etapa final, que visa responder às perguntas analíticas definidas na primeira etapa do trabalho.

## 2. Análise das Fontes de Dados

Nesta secção apresentamos todos as fontes de dados (adicionais e disponibilizadas), bem como a identificação de erros executada e sua correção e posterior análise dos dados. As fontes de dados disponíveis são: espectadores, canais vistos pelos espectadores, programação dos canais, tipos de programas, classes sociais dos espectadores e feriados e datas festivas.

### 2.1. Espectadores

O ficheiro “espetadores.csv” contém dados de espectadores. Cada registo contém oito campos separados por vírgulas, como por exemplo:

6, 3001, "Gr. Lisboa", "Femin.", "DDC", "+64", "D", #1996-01-01#

O significado de cada um dos campos de um registo, segundo a ordem em que aparecem, é explicado na Tabela 1.

Tabela 1 - Significado de cada um dos campos de um registo de espectadores, segundo a ordem em que aparecem

#	Campo	Tipo de Dados	Descrição	Exemplo
1	ID	Número inteiro	Identificador único de registo	6
2	Código	Número inteiro	Identificador único de espectador	3001
3	Região	Texto	Região do país da residência do espectador	"Gr. Lisboa"
4	Sexo	Texto	Masculino ou feminino	"Femin."
5	DonaDeCasa	Texto	Se o espectador trabalha em casa ou não	"DDC"
6	EscalãoEtário	Texto	Escalão etário do espectador	"64"
7	Classe	Texto	Classe social do espectador	"D"
8	Data	Data	Data de criação do registo	#1996-01-01#

#### 2.1.1. Tratamento de Dados

Para analisar o ficheiro “espetadores.csv” foi utilizado o software RStudio, aliado a um script R fornecido inicialmente pelo docente, tendo sido por nós adaptado posteriormente.

Através da linha de código “`print(NA_values <- sapply(espetadores, function(x) which(is.na(x))))`” não foram encontrados “missing values”. Esta informação foi contestada pela função “summary”, que no campo “região” encontrou duas entradas de “Região – Z”. Estas linhas foram eliminadas manualmente.

Na Tabela 2, Gráfico 1,2,3 e 4, encontram-se os dados de análise estatística a parte das variáveis em estudo, já com os dados corrigidos.

Tabela 2 - Análise de estatística descritiva sobre os campos "id", "codigo" e "data" da variável "espetadores" obtida através da função "summary"

	id	codigo	data
Min.:	1	240	1996-01-01
Mean:	152177	20892256	1996-03-31

Max.:	304353	34105603	1996-06-30
-------	--------	----------	------------

Tabela 3 - Análise de estatística descritiva sobre os campos "regiao", "sexo", "donadecasa", "escalaoetario" e "classe" da variável "espetadores" obtida através da função "summary"

regiao		sexo		dona de casa		escalao etario		classe	
Gr. Lisboa:	70597	Femin.:	111957	DDC:	145020	4-14:	29865	A/B:	67314
Gr. Porto:	50508	Masc.:	192394	nDDC:	159331	15-24:	50277	C1/C2:	88287
Interior:	39156					25-34:	38118	D:	148750
Lit. Norte:	42875					35-44:	46587		
Lit. Centro:	43808					45-54:	41351		
Sul:	57407					55-64:	47719		
						+64:	50404		

Para além dos erros referidos acima, também foram encontrados alguns **dados incoerentes**, nomeadamente no que toca a espectadores que mudam de atributos (como por exemplo de sexo). Estas linhas foram eliminadas.

Por exemplo: espetadores com o mesmo código apresenta valores discordantes para os atributos género e dona de casa.

153943	240	Lit. Centro	Masc.	DDC	35-44	C1/C2	#1996-04-01#
167460	240	Lit. Centro	Femin.	nDDC	35-44	C1/C2	#1996-04-09#

Imagem 1 - Exemplo de linhas com erros detetados

Os dados foram analisados visualmente com recurso à ferramenta Power BI, da Microsoft, na qual foram obtidos os gráficos 1, 2, 3 e 4. Para efeitos de “número de espectadores” foram tidos em consideração o número de registos individuais.

Pode verificar-se pelos gráficos que há uma maioria de registos do sexo masculino, de classe social D (trabalhadora) e da zona da Grande Lisboa. Há também uma ligeira predominância de pessoas da faixa etária “+64”.

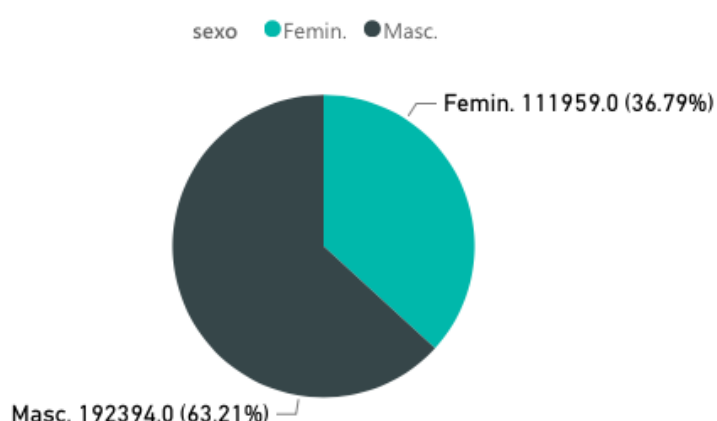
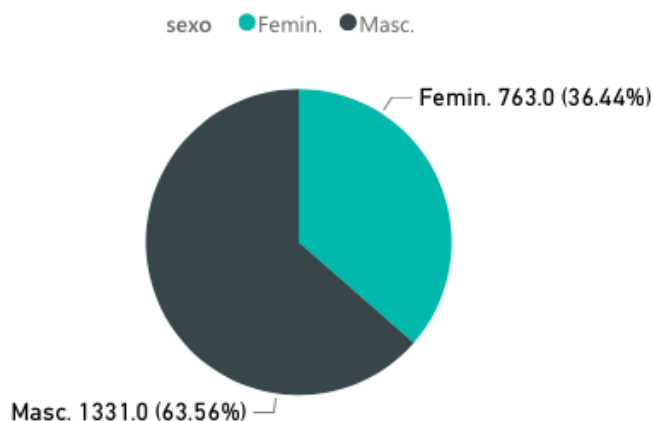


Gráfico 1 - Contagem e percentagem de códigos de espetadores e registos por sexo

O gráfico à esquerda diz respeito à totalidade de espetadores (2071), enquanto o segundo não tem em conta esta totalidade, mas sim a quantidade de registos, ou seja, o facto de um mesmo espetador com um certo código poder ter “id”s diferentes. Observa-se que a proporção é muito semelhante.

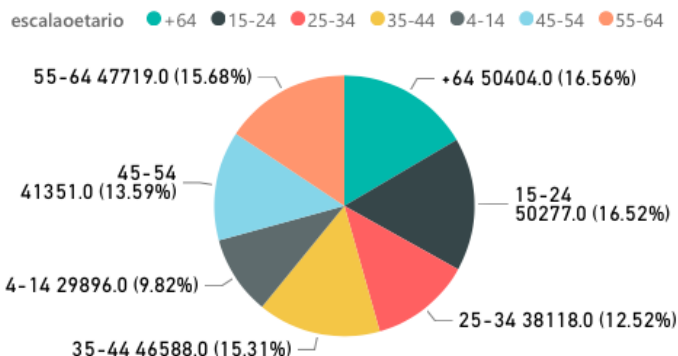
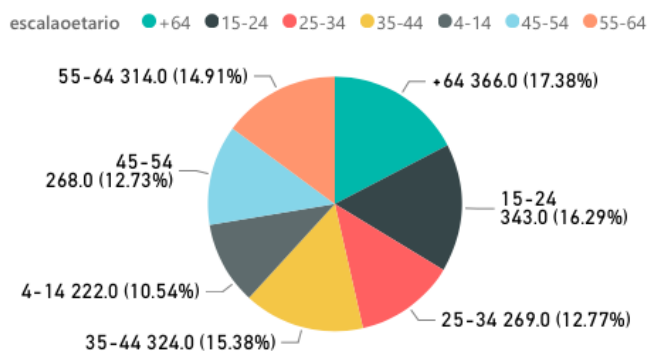


Gráfico 2 - Contagem e percentagem de códigos de espetadores e registos por escalão etário

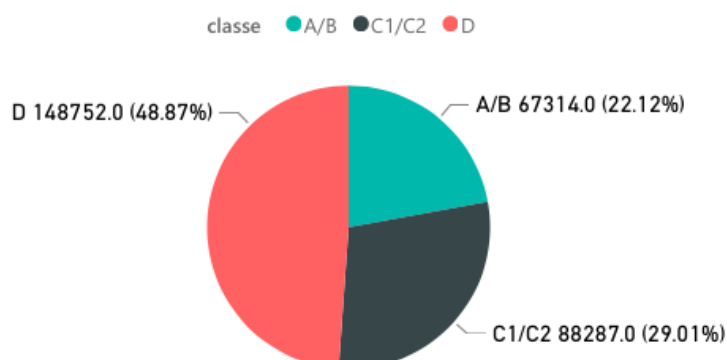
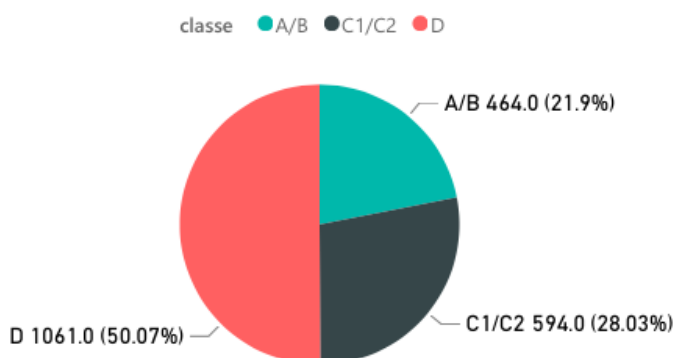


Gráfico 3 - Contagem e percentagem de códigos de espetadores e registos por classe social

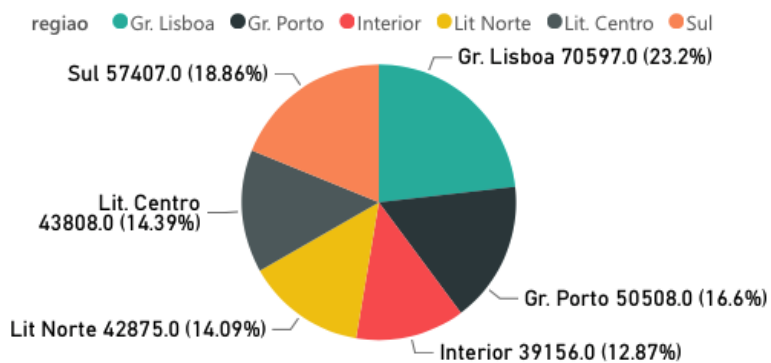
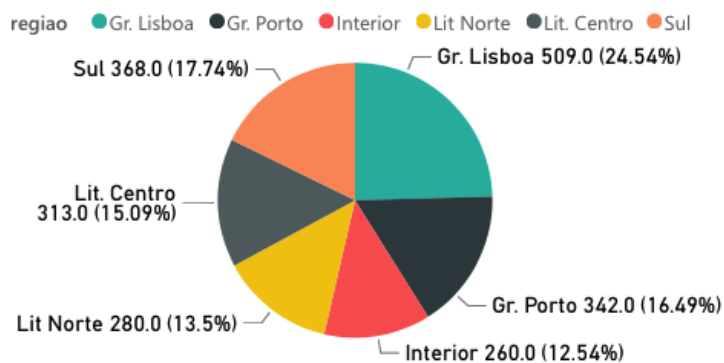


Gráfico 4 - Contagem e percentagem de códigos de espetadores e registos por região

Tal como no gráfico 1, os gráficos 2, 3 e 4 encontram-se separados: à esquerda é apresentado o gráfico que respeito à totalidade de espetadores (2071), enquanto o gráfico da direita se refere à quantidade de registos únicos. Novamente, a proporção entre os dois gráficos mantém-se semelhante.

## 2.2. Tipos de Programas Televisivos

O ficheiro “tipologia.tsv” guarda uma classificação dos vários tipos de programas televisivos.

A interpretação de cada um dos dois campos que formam um registo é esclarecida na Tabela 4, e uma listagem completa dos tipos de programa principais é apresentada na Tabela 5.

O tipo de programa é representado por uma sequência de até três letras (por exemplo abc), em que a primeira letra representa o tipo mais genérico do programa (ie, o tipo principal); a segunda letra indica o subtipo de programa; a terceira letra indica o último nível hierárquico do tipo de programa.

*Tabela 4 - Significado de cada um dos campos de um registo de programas televisivos, segundo a ordem em que aparecem*

#	Campo	Tipo de Dados	Descrição	Exemplo
1	Tipo	Texto	Identificador hierárquico do tipo de programa	abc
2	Designação	Texto	Designação do nível hierárquico do tipo de programa	Desenho Animado

*Tabela 5 - Designação dos Tipos Principais de Programas*

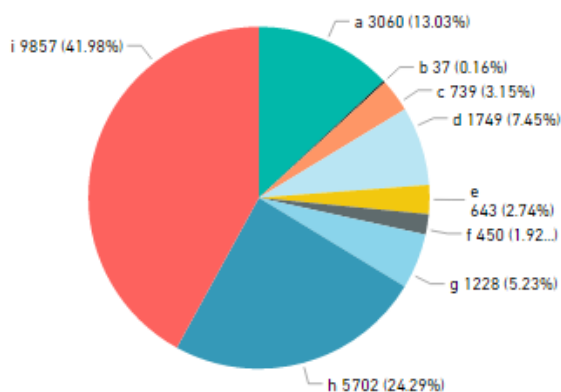
Identificador	Tipo Principal
a	Ficção
b	Eruditos
c	Variedades/divertim.
d	Informação
e	Cultura/conhecimento
f	Desporto
g	Juventude
h	Publicidade
i	Diversos
z	Outros

Para uma melhor compreensão dos dados, os mesmos foram analisados visualmente tendo sido obtido o conjunto de gráficos 5. A contagem total dos tipos de programa foi obtida através da leitura exclusiva da primeira letra do campo “tipo” dos dados referentes à programação.



Contagem dos Tipos Principais de Programas

Tipos de Programas a b c d e f g h i z



Tipos de Programa Principal por Canal

Tipos de Programas a b c d e f g h i z

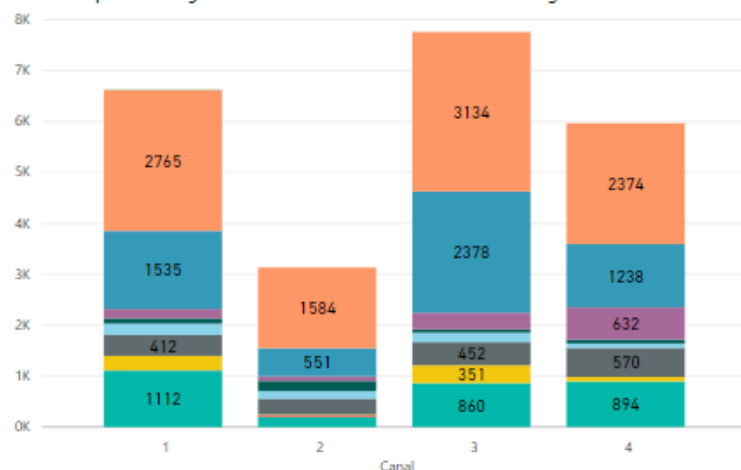


Gráfico 5 - Contagem dos Tipos Principais de Programas e Tipos de Programas por Canal

Através destes gráficos pode verificar-se que a maioria de visualizações é referente a programas do tipo “diversos”, sendo as duas categorias seguintes mais vistas as referentes a “publicidade” e “ficção”.

Pode ainda observar-se que as proporções dos tipos de programas se mantêm semelhantes nos vários canais.

### 2.3. Canais Vistos pelos Espectadores

O ficheiro “audiencias.csv” contém registos dos tempos de visualização de canais por cada espetador, estando os vários campos separados por vírgulas, como no exemplo seguinte:

57, #1996-01-01#, 1, 6, #1996-01-01 14:47:00#, #1996-01-01 14:53:00#

O significado de cada um dos seis campos que formam um registo, pela ordem em que surgem, é apresentado na Tabela 5.

Tabela 6 - Significado de cada um dos campos de um registo dos canais observados, segundo a ordem em que aparecem

#	Campo	Tipo de Dados	Descrição	Exemplo
1	ID	Número inteiro	Identificador de registo do espetador	57
2	Data	Data	Data de criação do registo	1996-01-01
3	Canal	Número inteiro	Número do canal visto pelo espetador	1
4	Duração	Número inteiro	Tempo de visualização do canal, em minutos	6
5	HoraInício	Data	Hora de início da visualização do canal	1996-01-01 14:47:00
6	HoraFim	Data	Hora de fim da visualização do canal	1996-01-01 14:53:00

De notar que os valores dos identificadores de registos de espetadores (primeiro campo, ID) estão relacionados com o campo ID do ficheiro “espetadores.csv”, descrito no capítulo 2.1. Assim, o valor “57” mencionado nesta secção é o mesmo que consta na secção sobre os espetadores de televisão.

#### 2.3.1. Tratamento de Dados

Através da linha de código “`print(NA_values <- sapply(audiencias, function(x) which(is.na(x))))`” obtivemos todos os índices correspondentes a variáveis com um **output** “NA”. Estes valores não definidos referem-se à não presença de horas no valor da data, tal como foi confirmado no Excel.

A informação obtida pela linha suprarreferida é corroborada pela linha de código “`summary(audiencias);`”, que devolve as informações presentes na Tabela 6.

Tabela 7 - Análise de estatística descritiva sobre os campos da variável “audiencias”

	id	data	duracao	horainicio	horafim		Canal
Min.:	57	1996-01-01	0.00	1996-01-01 02:00:00	1996-01-01 02:01:00	“1”:	695774
Mean:	149897	1996-03-29	24.93	1996-03-30 14:39:29	1996-03-30 15:04:59	“2”:	268942
Max.:	304353	1996-06-30	1055	1996-07-01 01:54:00	1996-07-01 01:55:00	“3”:	772158
						“4”:	411176

Através do software Excel foi feita uma ordenação das colunas “horainicio” e “horafim” (à vez) de acordo com o número de caracteres para isolar as entradas que não se encontravam no formato “AAAA-MM-DD HH:MM:SS”.

Verificou-se que as entradas neste formato correspondiam à hora “00:00:00”, pelo que 3221 entradas da coluna “horainicio” e 3858 entradas da coluna “horafim” foram corrigidas para conter a hora “00:00:00”.

Tal como para os dados do capítulo 2.1, também aqui foram elaborados gráficos para uma melhor interpretação dos dados.

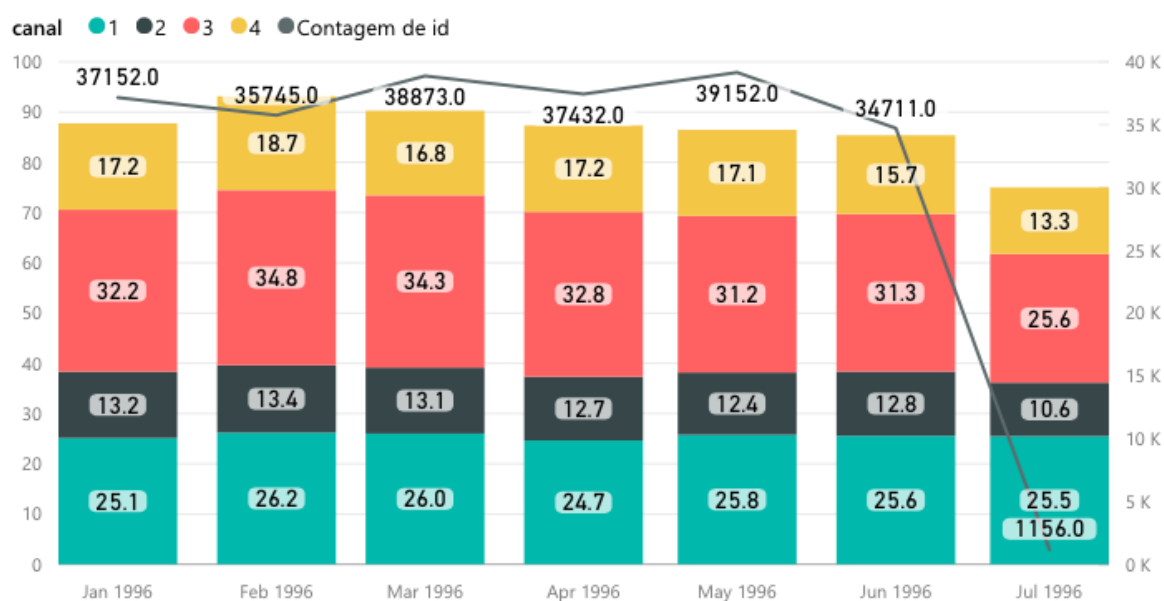


Gráfico 6 - Evolução da duração média, em minutos, por canal e mês

No gráfico 6 pode observar-se a evolução da duração média de visualização de cada canal, ao longo dos vários meses do semestre em análise. Em todos os meses, a maioria do *share*<sup>1</sup> vai para o canal 3.

De notar uma certa homogeneidade que apenas é quebrada em julho, devida ao facto de que neste mês apenas registadas 1156 visualizações pertencentes às primeiras horas da madrugada do dia 1 de julho.

<sup>1</sup> Percentagem de audiência de um canal/programa relativamente à audiência do total de televisão, para o mesmo período.

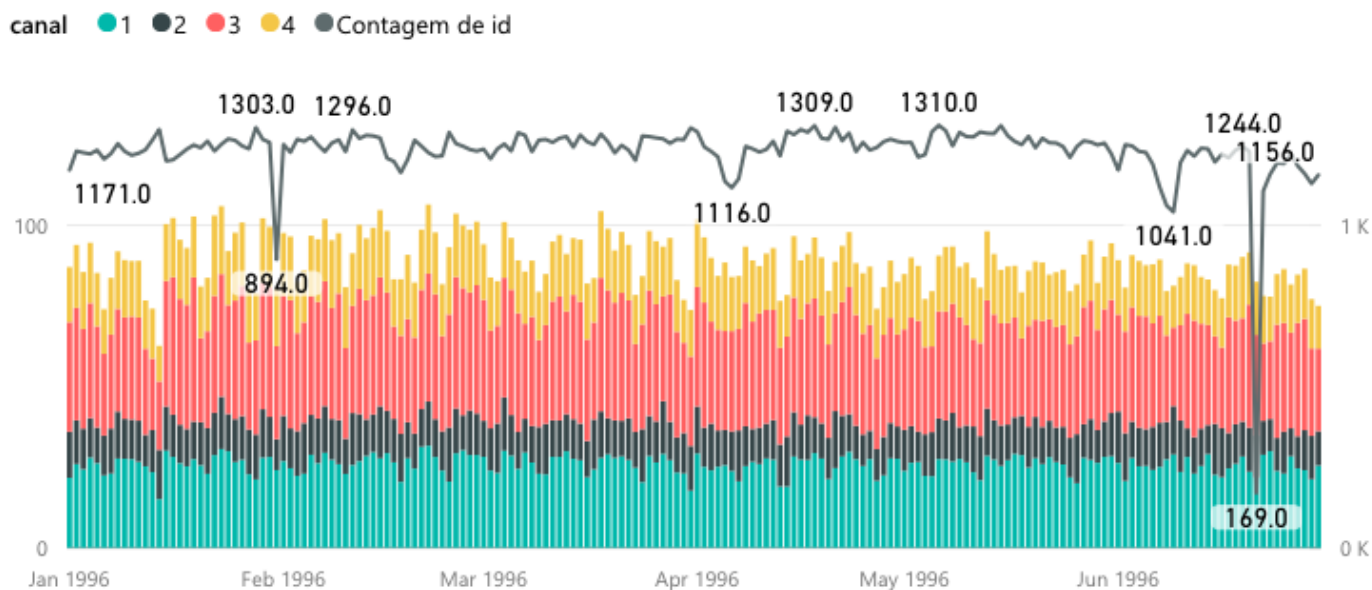


Gráfico 7 - Evolução da duração média, em minutos, por canal e dia (1 jan. a 30 junho)

No gráfico 7 apresenta-se a evolução diária dos registos de visualizações, assim como a evolução das durações por canal. Continua a verificar-se, assim como no gráfico 4, que a maioria do share vai para o canal 3.

Observam-se 2 reduções significativas no número de contagens de espetadores: a primeira no dia 31 de janeiro de 1996 (onde se verificaram apenas 894 registos), e a segunda no dia 21 de junho (onde se verificaram apenas 169 registos).

No gráfico 8 é visível a duração média (em minutos) segundo os dias da semana a que os registos se referem, e as faixas etárias dos respetivos espetadores.

De registar que os dias da semana com maior número de espetadores (em média) são terça e quarta-feira, sendo que os dias onde seria de esperar um maior número de espetadores (sábado e domingo) se encontram a meio e no fim da lista, respetivamente.

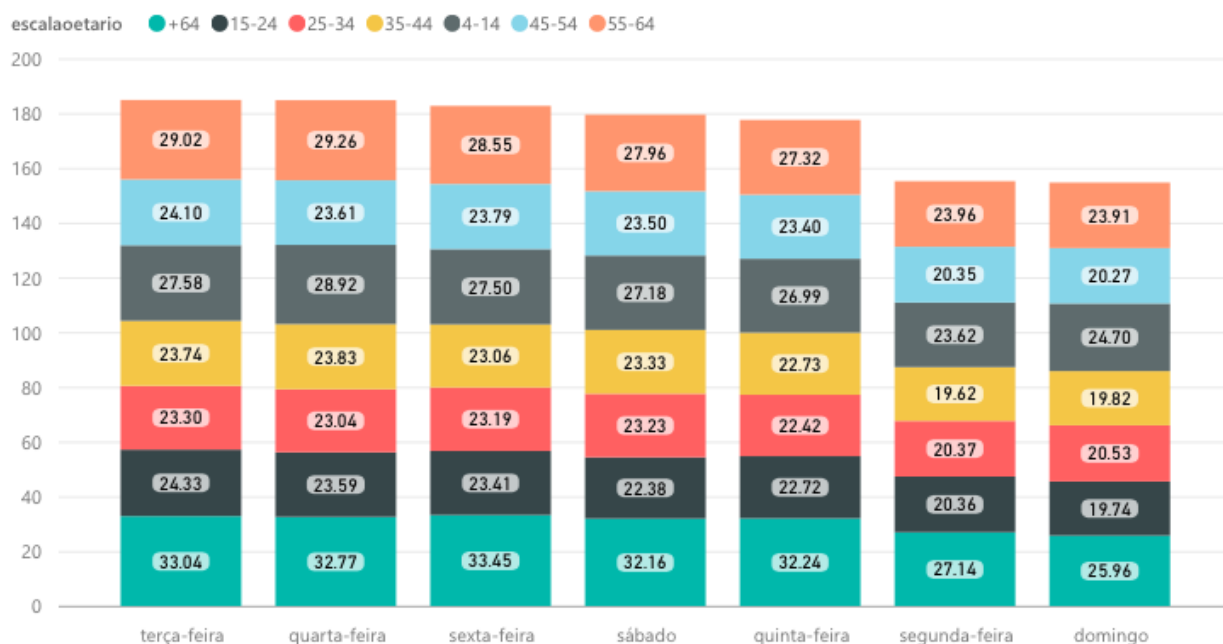


Gráfico 8 - Duração média, em minutos, por dia da semana e escala etária, ordenada segundo a grandeza total

Por fim, no gráfico 9 é visível a duração média de visualização ao longo dos dias da semana, por cada classe social. Tal como no gráfico 3, verifica-se uma maioria da classe D (trabalhadora), seguida da classe C1/C2 (média/média baixa).

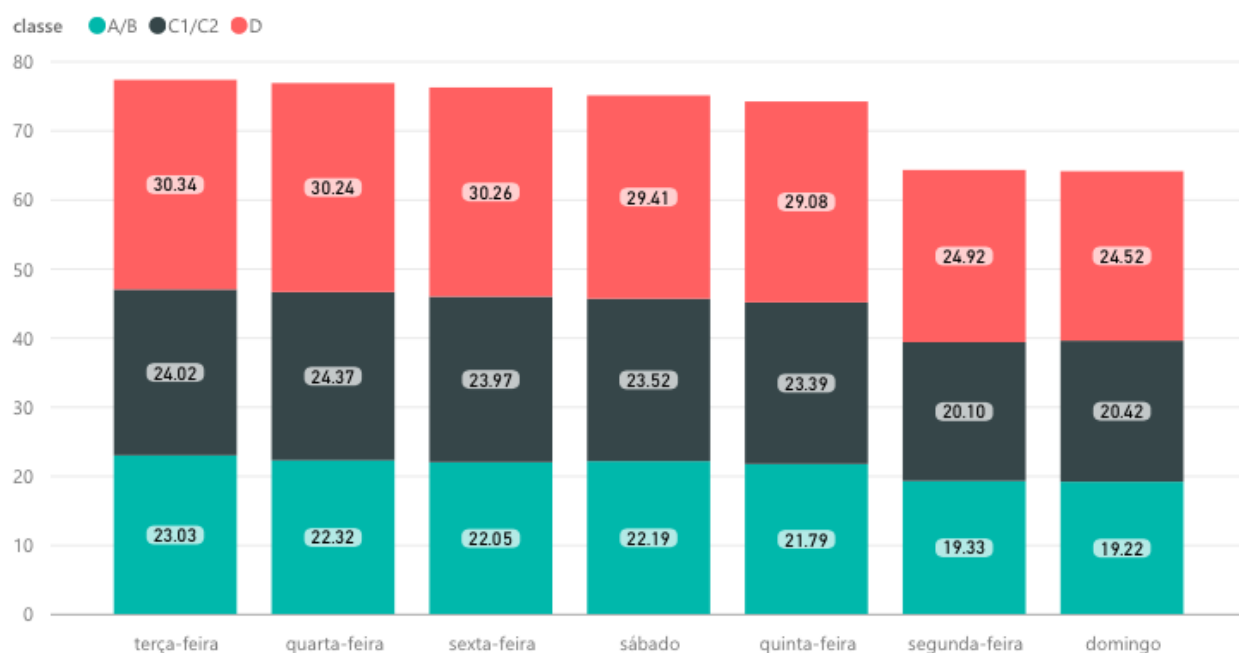


Gráfico 9 - Duração média, em minutos, por dia da semana e classe social, ordenada segundo a grandeza total

## 2.4. Programação dos Canais Televisivos

Esta fonte de dados é constituída por vários ficheiros, cada um guardando a programação televisiva tal qual foi emitida num determinado dia. Estão disponíveis tantos ficheiros quantos os dias do primeiro semestre de 1996, com nomes que identificam univocamente o dia a que respeitam.

A extensão PET vem de origem, mas o conteúdo é semelhante ao de um ficheiro com campos separados por vírgulas (.csv), com a diferença de cada linha terminar com o símbolo “;”.

1, 20000, 2775, 0, "SESSAO DUPLA I", "CLASSE", "P", "aae", 0;

O significado de cada um dos nove campos de um registo é ilustrado na Tabela 8.

Tabela 8 - Significado de cada um dos campos de um registo de programação, segundo a ordem em que aparecem

#	Campo	Tipo de Dados	Descrição	Exemplo
1	Canal	Número inteiro	Número do canal no ar	1
2	Horainício	Número inteiro	Hora inicial do programa, no formato hhmmss	20000
3	Duração	Número inteiro	Duração do conteúdo televisivo, em segundos	2775
4	Zero	Número inteiro	Sem significado	0
5	Nome1	Texto	Nome do conteúdo televisivo	"SESSAO DUPLA I"
6	Nome2	Texto	Um segundo nome do conteúdo televisivo	"CLASSE"
7	Classificação	Texto	Classificação do conteúdo, detalhada a seguir	"P"
8	Tipo	Texto	Tipo do conteúdo, de acordo com a tipologia em cima	"aae"
9	ParteTodo	Número inteiro	Se representa o conteúdo todo ou uma das suas partes	0

Todos os ficheiros PET guardam registos com hora de início às 2h00 da noite (valor 20000 em Horainício) e registam a sequência de conteúdos televisivos emitida num período de 24 horas; por exemplo, o valor 253015 representa "25" horas, 30 minutos, e 15 segundos, ou seja, cerca da uma e meia da noite do dia seguinte.

Relativamente ao campo “Classificação”, este pode tomar três valores distintos: "P" para programa, "B" para intervalo comercial, e "I" para publicidade ao próprio canal.

Por fim, o campo “ParteTodo” indica se o registo diz respeito a um programa como um todo (valor 0) ou a uma das partes (valor 1), sendo que um valor 0 pode incluir intervalos.

Em alguns ficheiros PET existem registos cujo valor da variável “ParteTodo” é 2, o que significa que se trata de uma parte de um todo com código 1, em vez de 0, isto é uma parte integrada dentro de outra parte. Nestas circunstâncias, a duração total do programa (com ParteTodo = 0) corresponde à soma das durações dos subprogramas

(com ParteTodo = 1), e a duração destes será a soma das dos sub-subprogramas (com ParteTodo = 2).

#### 2.4.1. Tratamento de Dados

Tal como para as fontes de dados referentes aos espetadores e canais vistos por estes, também aqui foi utilizado um script R para auxílio à análise de estatística descritiva dos dados em estudo.

Uma vez que algumas das variáveis da presente fonte de dados possuem valores descritivos em grande quantidade, apenas algumas variáveis são apresentadas na Tabela 9, de acordo com a função “summary(todos.dados.pet);”.

Tendo em conta que a coluna “Zero” apresentada nos dados originais não tinha significado, e para uma melhor visualização e entendimento dos dados, esta foi eliminada.

*Tabela 9 - Análise de estatística descritiva sobre os campos "Horalnicio", "Duracao", "Canal", "Classificacao" e "ParteTodo" da variável "todos.dados.pet" obtida através da função "summary"*

	Horalnicio	Duracao	Canal	Classificacao	ParteTodo			
Min.:	20000	2.0	1	6625	B	6221	0	9543
Mean:	168543	573.8	2	3133	I	9121	1	12895
Max.:	255957	22478.0	3	7757	P	8137	2	1041
			4	5964				

Usando um script R para ordenar os dados por valores e procura por valores NA, foram encontrados os seguintes erros:

##### 1. Vírgulas em falta

Exemplo: Ficheiro “19960607.pet”, linha 399:

`3, 22934, 5, 0, "PATROCINIO", "1""B", "hc", 1;`

Correção: vírgula adicionada manualmente.

##### 2. Aspas em falta

Exemplo: Ficheiro “19960217.pet”, linha 384:

`3, 90044, 97, 0, "INT.APRES.PROGRAMAS",", "I", "ib", 0;`

Correção: aspa adicionada manualmente, mas quando o atributo não tem valor (como no exemplo acima) os atributos foram ignorados.

##### 3. Valores de atributos em falta

Exemplo 1: Ficheiro “19960605.pet”, linha 404:

`, 43030, 2576, 0, "OUTRO CINEMA", "KRUSH GROOVE", "P", "aak", 1;`

Correção: falta o número do canal, portanto as linhas com este tipo de erro foram eliminadas uma vez que o canal não pode ser inferido por outros dados.

Exemplo 2: Ficheiro “19960508.pet”, linha 396:

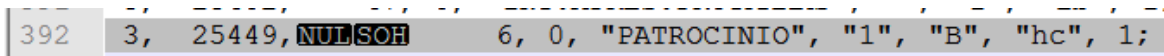
```
3, 24202, 43, 0, "PUBLICIDADE ASSOCIADA", "1", "B", "ha",
```

Correção: falta o atributo “ParteTodo”, pelo que as linhas detetadas com este tipo de erro foram eliminadas.

#### 4. Caracteres inválidos

Exemplo: Ficheiro “19960610.pet”, linha 392:

```
3, 25449, 6, 0, "PATROCINIO", "1", "B", "hc", 1;
```



```
392 3, 25449, NULSOH 6, 0, "PATROCINIO", "1", "B", "hc", 1;
```

Imagem 2 - Representação gráfica do exemplo supracitado para exemplificação dos caracteres inválidos

Correção: os caracteres inválidos foram removidos, tendo os restantes valores sido preservados.

Por fim, foi criado um ficheiro “programas.tsv”, com todos os dados extraídos dos ficheiros “.pet”.



## 2.5. Classes Sociais dos Espectadores

O ficheiro “classes.tsv” descreve o significado das letras A, B, ..., que identificam classes sociais.

O significado de cada um dos campos de um registo é explicado na Tabela 10, e a lista completa das diferentes classes sociais e sua descrição é apresentada na Tabela 11.

*Tabela 10 - Significado de cada um dos campos de um registo das classes sociais dos espetadores, segundo a ordem em que aparecem*

#	Campo	Tipo de Dados	Descrição	Exemplo
1	Classe	Texto	Classe social	A
2	Estatuto	Texto	Estatuto social	Classe média/alta
3	Ocupação	Texto	Ocupações representativas	Gestor, administrador, ou profissional de topo

*Tabela 11 - Listagem das classes sociais e respetivos estatutos e ocupações*

Classe	Estatuto	Ocupação
A	Classe média/alta	Gestor, administrador, ou profissional de topo
B	Classe média	Gestor, administrador, ou profissional intermédio
C1	Classe média/baixa	Supervisor ou empregado de escritório, gestor, administrador, ou profissional júnior
C2	Classe trabalhadora qualificada	Trabalhador manual qualificado
D	Classe trabalhadora	Trabalhador manual pouco ou não qualificado
E	Aqueles com menor nível de subsistência	Pensionistas sem outros rendimentos, trabalhadores temporários

De notar que o campo “Classe” é também usado no ficheiro “espetadores.csv” (ver capítulo 2.1), embora neste caso existam várias ocorrências em que num mesmo valor são concatenados dois identificadores de classe (por exemplo, A/B).

## 2.6. Fonte de Dados Adicionais

Foi criado um ficheiro de extensão “tsv” (separado por tabulações), que contém feriados e datas festivas de 1996, com base na informação disponível na Internet<sup>2</sup>.

*Tabela 12 - Listagem da fonte de dados adicional “calendário”*

#	Campo	Tipo de Dados	Descrição	Exemplo
1	Data	Data	Dia do feriado	01/05/1996
2	Feriado	Texto	Indica se é feriado ou não	Feriado
3	Descrição	Texto	Nome do feriado e nomes alternativos	Dia do Trabalhador

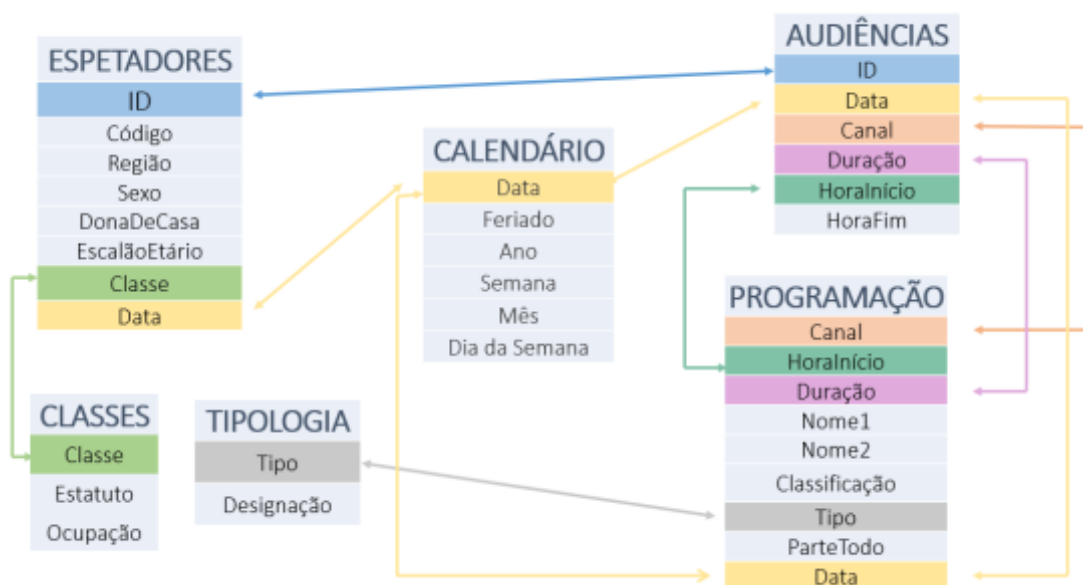
No diagrama de dados existe uma fonte de dados “Calendário” que apresenta outros atributos como ano, semana, mês e dia da semana, cuja representação neste ficheiro

<sup>2</sup> <https://www.calendarr.com/portugal/calendario-1996/>

não é necessária, visto que os seus respetivos valores podem ser obtidos recorrendo a funções em SQL ou Excel.

### 3. Relação entre Fontes de Dados

Para uma melhor compreensão dos dados disponibilizados, foi realizado um diagrama que relaciona as diferentes fontes de dados, direta ou indiretamente.



Esquema 1 - Diagrama das Fontes de Dados e suas relações.

Analisando o resultado final, podemos retirar algumas informações:

- O ficheiro referente às audiências contém os dados de cada registo feito pelo espetador, que é o mesmo registado no ficheiro espetadores. Para além disto, contém a data desse registo que também pode se obter no ficheiro relativo aos espetadores;
- No ficheiro alusivo audiências, um dos parâmetros é o canal, ou seja, podemos ter acesso ao que um espetador viu cruzando os dados com o ficheiro referente à programação e posteriormente saber o tipo de programa através do ficheiro que se refere às diferentes tipologias;
- Finalmente, para saber a classe social de cada espetador pode-se fazê-lo cruzando o ficheiro alusivo aos espetadores com o ficheiro referente às classes.

Uma fonte de dados extra foi acrescentada, “Calendário”, que nos permite aceder aos feriados e datas comemorativas do primeiro semestre de 1996. Estes dados são cruzados com o ficheiro das audiências e da programação.

É possível observar um exemplo destas relações no esquema 3 (na seção Anexos), onde carregando na seção “Femin.” do gráfico circular, no canto superior direito, se verifica a resposta das variáveis com as quais interage nos restantes gráficos.

A resposta de maior destaque, neste caso, será uma clara superioridade da duração média de visualização por raparigas da faixa etária dos 4 aos 14 anos, à sexta feira, com

uma média de duração de 94 minutos, conforme se pode verificar no gráfico do canto superior esquerdo, a cinzento.

## 4. Modelação Dimensional

### 4.1. Processo de Negócio

O consumo de televisão varia ao longo do ano. Um dos fatores diretamente relacionado com essa variação do consumo é a sazonalidade (ao longo do ano). Entender os espetadores que contactam com este meio em diferentes períodos do ano e também do dia são avaliações importantes para a compreensão da evolução do consumo televisivo.

Nas estações televisivas é necessário analisar os hábitos dos telespetadores para adaptar os conteúdos transmitidos. Assim, baseando-nos na amostra disponibilizada, o processo de negócio em foco será as **tendências das audiências televisivas tendo em conta um determinado período de tempo**, nomeadamente durante o primeiro semestre de 1996.

A análise deste processo de negócio consistirá nomeadamente em estudar o que acontece aos diferentes parâmetros, por exemplo faixa e classe etária, região, duração da visualização.

### 4.2. Perguntas Analíticas

Para facilitar este estudo, elaborámos algumas questões específicas para este processo:

- Quais são os programas televisivos mais vistos por cada faixa etária e classe social ao longo da semana e durante o fim-de-semana?
- Qual a média de espetadores por canal em cada dia durante o primeiro semestre de 1996?
- Quanto tempo cada espetador vê televisão durante a semana e ao fim-de-semana?
- Por cada região do país, quem (grupo social e se trabalha ou não em casa) passa mais e menos tempo a ver televisão tendo em conta a ocupação?
- Tendo em conta os feriados nacionais e datas comemorativas no primeiro semestre de 1996, qual a média de espetadores por faixa etária e classe social?

### 4.3. Definição do Grão

Tendo em conta o processo de negócio em estudo, referente às tendências das audiências televisivas para um determinado período de tempo, é necessário registar eventos detalhado, tendo sido definido o seguinte grão para a tabela de factos:

*“Um espetador vê um programa, numa determinada data e hora, durante um período de tempo.”*

De notar que não há distinção entre as partes de um programa, sendo apenas analisada a duração do programa que o espetador vê.

Factos com este grão podem ser utilizados para responder a questões como a seguinte:

*“Quais os programas mais vistos entre as 15h e 16h?”*

Por sua vez, uma resposta hipotética poderia ser:

*“O programa do tipo “c” (“variedades/diver.”) do canal 4, visto por x espetadores, com uma duração média de y minutos.”*

Neste caso, o **tipo de tabela de factos** utilizado é do tipo **transaccional**, ou seja, a tabela que registam eventos que ocorrem em determinados momentos – cada facto aconteceu num ponto no tempo, em que cada linha corresponde ao registo de um novo evento conforme o grão definido.

Recorrendo ao nosso grão, cada linha da tabela de factos irá registar o que um espetador visualizou num dado momento ou período de tempo.

#### 4.4. Dimensões do Negócio

De acordo com os atributos definidos para a tabela de factos, foram identificadas quatro dimensões: **Espetador**, **Programa**, **Data** e **Horário**.

##### 4.4.1. Dimensão Espetador

A tabela Dimensão Espetador gera tantas linhas quanto o número de espetadores (isto é, tantas linhas quanto o número de códigos diferentes) existentes. Neste caso, a tabela de dimensão terá 2071 linhas.

Foi identificada uma hierarquia correspondente ao estatuto/classe social, que é definida da seguinte forma:

- Estatuto 1 (Estatuto Principal)
  - Estatuto 2 (Estatuto Secundário)

O significado de cada atributo da tabela é sumariado na tabela 13.

*Tabela 13 - Significado de cada atributo da tabela “Dimensão Espetador”*

Atributos	Tipo de Dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta Espetador	Número inteiro	Identificador único de espetador	Criado manualmente	1
Código (Chave Natural)	Número inteiro	Código identificador do espetador	Espetadores	6
Género	Texto	Masculino ou feminino		"Femin."
Escalão Etário	Texto	Escalão etário do espetador		" +64"
Região	Texto	Região do país da residência do espetador		"Gr. Lisboa"
Estatuto 1	Texto	Classe social principal do espetador		"Classe média/alta"

Ocupação 1	Texto	Descrição da ocupação do espetador, tendo em conta o seu estatuto		"Gestor, administrador, ou profissional de topo"
Estatuto 2	Texto	Classe social secundária do espetador, caso existente		"Estatuto não definido"
Ocupação 2	Texto	Descrição da ocupação do espetador, tendo em conta o seu estatuto, caso existente		"Ocupação não definida"
Dona de Casa	Texto	Se o espetador trabalha em casa ou não		"DDC"

#### 4.4.2. Dimensão Programa

A tabela referente à Dimensão Programa gera tantas linhas quantos os programas registados, sendo que foi considerado "um programa" como a junção de todos os múltiplos registos de um mesmo programa, ou seja, consideraram-se os programas cuja "parte todo" era igual a 0.

Foram também considerados apenas os programas de classificação "P", ou seja, os intervalos comerciais ("B") e publicidades ao próprio canal ("I") foram ignoradas, por não serem de relevo para o processo de negócio em análise.

Nesta dimensão identificaram-se 2 hierarquias distintas: uma para os nomes dos programas, e outra para o tipo de programa. Estas hierarquias são demonstradas em seguida, e estão incluídas na descrição dos elementos da dimensão, na tabela 14.

##### → Nome Programa

- Nome geral
  - Nome específico

##### → Tipo Programa

- Tipo
  - Categoria
    - Género

Tabela 14 - Significado de cada atributo da tabela "Dimensão Programa"

Atributos	Tipo de Dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta Programa	Número inteiro	Identificador único do programa	Criado manualmente	1
Nome Geral	Texto	Nome do conteúdo televisivo	Programação	"FILME"
Nome específico	Texto	Um segundo nome do conteúdo televisivo		"CURTO CIRCUITO II"
Canal	Texto	Número do canal no ar		4
Tipo	Texto	Tipo do conteúdo		"FICÇÃO"
Categoria	Texto	Tipo do conteúdo		"FILME"
Género	Texto	Tipo do conteúdo		"Comedia"

#### 4.4.3. Dimensão Data

A tabela “Dimensão Data” gera tantas linhas quanto o número de dias do período em análise; neste caso, serão 183 linhas.

Na dimensão Data identificaram-se 2 hierarquias possíveis a partir do ano, que são identificadas em seguida, assim como na tabela 15.

→ **Ano**

- Semana do ano
  - Dia da Semana
  - Fim-de-Semana
- Mês
  - Dia do mês
  - Data Comemorativa
    - Feriado

Tabela 15 - Significado dos atributos da tabela “Dimensão Data”

Atributos	Tipo de dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta de Data	Número inteiro	Identificador único da data	Criado manualmente	19960401
Data Completa (Chave Natural)	Data	Data completa		1996-04-01
Dia do Mês	Número inteiro	Número referente ao dia do mês	Calendário	4
Nome do Mês	Texto	Nome do mês do ano		“janeiro”
Ano	Número inteiro	Número referente ao ano em questão		1996
Dia da Semana	Texto	Nome do dia da semana		“Segunda-feira”
Semana do Ano	Número inteiro	Número referente à semana do ano, de 1 a 52		1
Fim-de-semana	Texto	Se o dia da semana é referente a fim-de-semana ou não		“dia de semana”
Indicador Data Comemorativa	Texto	Se o dia é uma data comemorativa ou não		“Data Não Comemorativa”
Nome Data Comemorativa	Texto	Nome da Data Comemorativa		“Data Não Comemorativa”
Indicador Feriado	Texto	Se o dia da semana é referente a um feriado ou não		“Não feriado”

#### 4.4.4. Dimensão Horário

Por forma a controlar o crescimento da dimensão data, a dimensão referente às horas do dia foi separada desta, evitando tornar a dimensão data numa dimensão monstra.

Esta tabela de dimensão gera tantas linhas quanto o número de horas, minutos e segundos de um dia, isto é, 86 400 linhas.

Para este caso, identificámos apenas uma hierarquia, referente ao período do dia. Novamente, esta hierarquia é demonstrada em seguida, assim como na tabela 16, referente aos atributos da dimensão

→ **Período do dia**

- Hora
  - Minutos
    - Segundos

Tabela 16 - Significado dos atributos da tabela “Dimensão Horário”

Atributos	Tipo de dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta Horário	Número inteiro	Identificador de Horário	Criado manualmente	124751
Horário Completo	Número inteiro	Identificador único do horário		“12:47:51”
Período do Dia	Texto	Se a hora do dia corresponde ao período da manhã (7h-12h), tarde (12h-18h), noite (18h-00) ou madrugada (00h-7h)		“Manhã”
Hora	Número inteiro	Hora do dia		12
Minutos	Número inteiro	Minutos correspondentes à hora do dia		47
Segundos	Número inteiro	Segundos correspondentes à hora do dia		51

#### 4.5. Registo de Mudanças Lentas

Para registar as mudanças de dados referentes aos espetadores, como por exemplo mudança de classe etária, classe social, região ou dona de casa, optámos pela técnica do Tipo 2 - **acrescentar uma linha** na tabela de Dimensão Espetadores. Esta estratégia é indicada para manter o histórico destas alterações.

Para além disto, foram acrescentadas 3 colunas: “Data Início”, “Data Fim” e “Em vigor”. Estas colunas permitem-nos organizar os dados de uma forma mais coerente e obter as respostas indicadas à pergunta analítica em questão.

Tabela 17 - Dimensão Espetador, tendo em consideração o registo de mudanças lentas.

Atributos	Tipo de Dados	Descrição	Origem dos Dados	Exemplo
Chave Substituta Espetador	Número inteiro	Identificador único de espetador	Criado manualmente	1
Chave Supernatural Espetador	Número Inteiro	Identificador único de espetador		
Código (Chave Natural)	Número inteiro	Código identificador do espetador	Espetadores	6
Género	Texto	Masculino ou feminino		"Femin."
Escalão Etário	Texto	Escalão etário do espetador		" +64"
Região	Texto	Região do país da residência do espetador		"Gr. Lisboa"
Estatuto 1	Texto	Classe social principal do espetador		"Classe média/alta"

Ocupação 1	Texto	Descrição da ocupação do espetador, tendo em conta o seu estatuto		"Gestor, administrador, ou profissional de topo"
Estatuto 2	Texto	Classe social secundária do espetador, caso existente		"Estatuto não definido"
Ocupação 2	Texto	Descrição da ocupação do espetador, tendo em conta o seu estatuto, caso existente		"Ocupação não definida"
Dona de Casa	Texto	Se o espetador trabalha em casa ou não		"DDC"
Data Início	Data	Data respeitante ao dia em que começou a ser visualizado um programa	Criado manualmente	1996-01-01
Data Fim	Data	Data respeitante ao dia em que terminou de ser visualizado um programa		1996-07-01
Em Vigor	Texto	Se a linha se encontra em vigor, <i>ie</i> , possui a informação atualizada/corrente		"TRUE"

#### 4.6. Medidas Numéricas

Para avaliar as tendências das audiências televisivas, foi adicionada uma medida aditiva "duração" à tabela de factos.

Esta medida refere-se ao período de tempo, em minutos e segundos, em que um espetador viu um programa, e permite o cálculo de valores agregados (ex. soma) ao longo de todas as dimensões.

O tempo mínimo considerado para a visualização de um canal é de 60 segundos, uma vez que esta é a frequência mínima de registo.



#### 4.7. Diagrama em Estrela do *Data Warehouse*

De acordo com o processo de negócio tido em conta, o respetivo grão e dimensões de negócio, foi construído o diagrama demonstrado em seguida, com as correspondentes tabelas de dimensões e de factos.

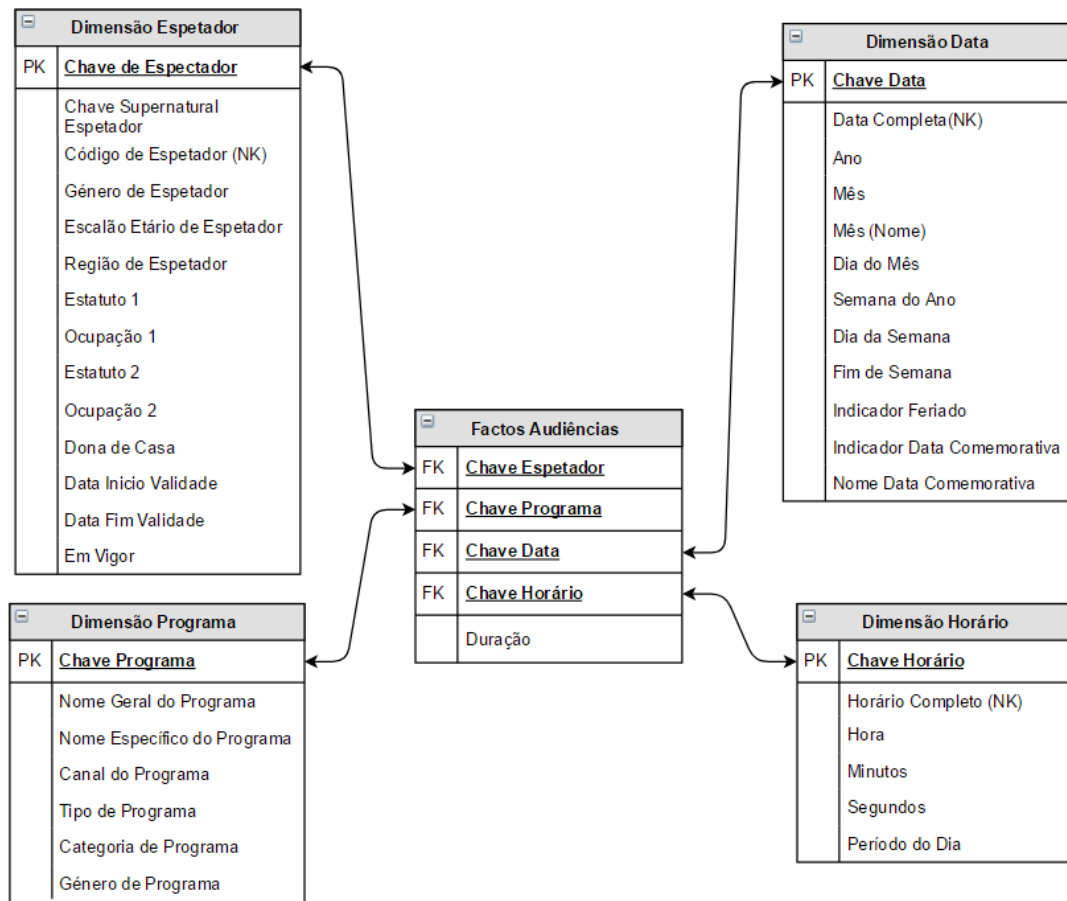


Diagrama 1 - Tabelas de Factos e Dimensões

## 5. Sistema ETL

Depois de definida a modelação dos dados e estruturado o *Data Warehouse*, é necessário povoá-lo com os dados organizados. Para tal, recorre-se ao sistema ETL. Este processo é uma das fases mais críticas na construção de um *data warehouse*, pois é nesta fase que grandes volumes de dados são processados.

**Extração, Transformação e Carga** (Extract, Transform & Load - ETL) são etapas de uma técnica que permite às organizações extrair dados de fontes de informação diversas e reformulá-los e carregá-los para uma nova aplicação (Data Warehouse, com recurso a base de dados) para análise.

Através do diagrama 2, pode obter-se uma visão geral de todas as fases seguidas neste processo.

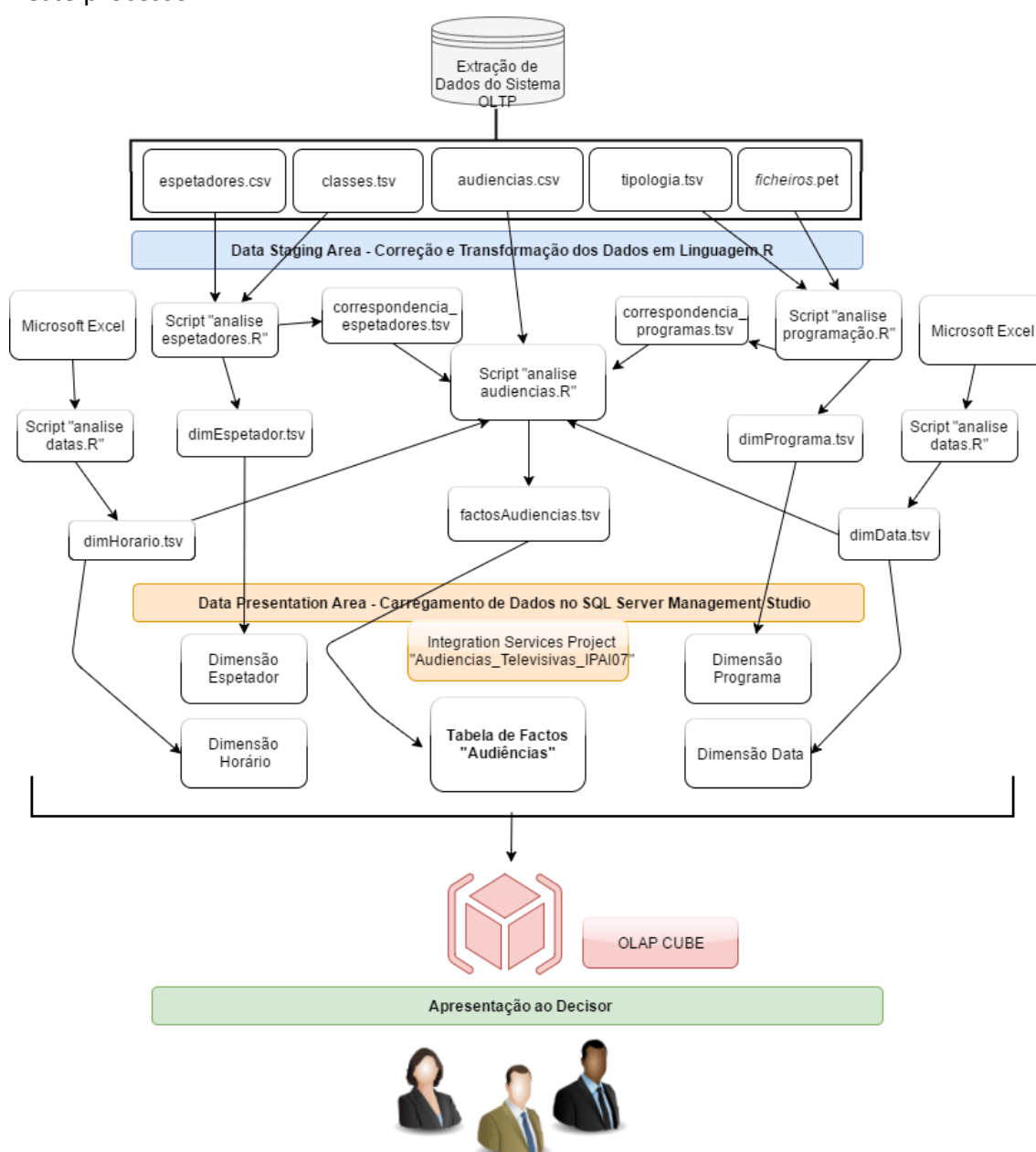


Diagrama 2 - Fluxo de dados e programas do sistema ETL

Os diversos programas elaborados e descritos no diagrama 2 são sumariados na tabela 18.

Tabela 18 - Responsabilidade da etapa "Extração" do sistema ETL e de cada programa

Programa	Responsabilidade	Entrada	Saída
analise datas.R	- Transforma a formação ficheiros de data e horários gerados em excel para dar o tipo correcto aos dados (nomeadamente remover 0 extra da chave).	data.tsv horario.tsv	dimData.tsv dimHorario.tsv
analise espetadores.R	- Substitui as abreviaturas pelas classes dos espetadores; - Corrige erros nos espetadores e remove dados sem sentido; - Processa mudanças lentas nos espetadores; - Torna os campos de espetador mais inteligíveis; - Cria a dimensão espetador; - Cria a tabela de correspondência entre espectadores extraídos e dimensão espectador.	espetadores.csv classes.tsv	dimEspecadores.tsv correspondencia_espetador.tsv
analise programas.R	- Substitui os tipos pela respetiva informação; - Seleciona os programas de interesse; - Cria a dimensão programa; - Cria a tabela de correspondência entre programas extraídos e dimensão programa.	Ficheiros “.pet” tipologia.tsv	dimPrograma.tsv correspondencia_programa.tsv
analise audiencias.R	- Cruza a informação registada nas audiências com programas e espectadores; - Gera a tabela de factos.	audiências.csv dimData.tsv dimHorario.tsv correspondencia_espetador.tsv correspondencia_programa.tsv	factosAudiencias.tsv

### 5.1. Extração dos Dados

A **extração** é a primeira etapa no processo de obtenção de dados para o ambiente do *data warehouse*. Extrair significa ler e compreender a fonte de dados, e transcrever os dados necessários para a *Data Staging Area*, para futura manipulação.

As fontes de dados originais para extração de dados (“classes.tsv”, “tipologia.tsv”, “audiências.csv”, “espetadores.csv” e ficheiros “.pet”) foram fornecidas pelo docente.

Para além destas, acrescentou-se uma fonte adicional “Calendário”. Estas informações de data e hora foram geradas no Microsoft Excel e guardadas em ficheiros: “data.tsv” e “horario.tsv”.

A estas fontes foi aplicada uma análise estatística (descrita na secção 2), que nos permitiu ter uma visão geral sobre como é que os dados se interligam e se comportam.

As diferentes ferramentas utilizadas para a extração dos dados e suas responsabilidades no processo são descritos na tabela 19.

Tabela 19 - Responsabilidades da etapa “Extração” do sistema ETL e de cada ferramenta

Ferramenta	Responsabilidade	Entrada	Saída
R Studio	→ <b>Receção</b> - Receção dos dados das várias fontes; - Deteção de alterações nos dados; - Aplicação de filtros para deteção; - Ordenação dos dados; - Primeira Limpeza dos dados; - Tratamento de exceções; - Eliminação de duplicados.	Audiencias.csv Espetadores.csv Ficheiros “.pet” Classes.tsv Tipologia.tsv	Audiencias.csv Espetadores.csv Programas.tsv
MS Excel	→ <b>Receção</b> - Criação manual de fonte de dados.	-	data.tsv horario.tsv

## 5.2. Transformação dos Dados

Assim que os dados são extraídos para a *Data Staging Area* (área de trabalho para processar dados em bruto), existem inúmeras potenciais transformações (mais complexas que numa primeira análise), como: tradução de valores codificados, aplicação da transformação apenas a determinadas categorias de linhas e/ou colunas, conflitos de domínio, lidar com elementos ausentes, fusão (*merging*) ou agregação dos dados.

Os dados-fonte “classes.tsv” e “tipologia.tsv” são meramente informativos e, portanto, não foram sujeitos a nenhuma correção; relativamente às restantes fontes de dados, os erros detetados e corrigidos são descritos nos respetivos subcapítulos da secção 2.

Os dados fornecidos foram corrigidos e preparados para construção das dimensões e tabela de factos definidas anteriormente, recorrendo a duas ferramentas: Microsoft Excel, e R Studio.

As responsabilidades de ambos os programas são descritas na tabela 20, e a transformação de cada dimensão é descrita em respetivo subcapítulo.

Tabela 20 - Responsabilidades da etapa “Transformação” do sistema ETL e de cada ferramenta

Ferramenta	Responsabilidade	Entrada	Saída
R Studio	→ <b>Integração</b> - Segunda Limpeza dos Dados;	Audiencias.csv Espetadores.csv	dimEspetador.csv

	- Geração de valores de chaves substitutas para todas as dimensões; - Fusão de duplicados; - Preenchimento de valores em falta; - Correspondência entre os diversos dados; - Criação das dimensões e tabela de factos.	Ficheiros “.pet”	dimPrograma.csv
MS Excel	→ <b>Integração</b> - Criação manual de dimensões.		data.tsv horario.tsv

#### 5.2.1. Dimensão Data e Dimensão Horário

As dimensões Data e Horário foram geradas automaticamente com recurso à ferramenta Microsoft Excel, sendo que cada célula corresponde a um dia (no caso da dimensão Data) e a uma hora, minuto e segundo (no caso da dimensão Horário).

Para a dimensão Data, para além das colunas referentes à data completa, dia, mês e ano, foram ainda adicionadas outras colunas descritivas, para uma melhor perceção por parte do decisor. São elas “nome do mês”, “semana do mês”, “dia da semana”, “fim de semana”, “indicador feriado”, “indicador data comemorativa”, “nome da data comemorativa”.

De igual forma, na dimensão Horário, para além das colunas referentes à hora completa, hora, minutos e segundos, foi adicionada uma coluna descritiva “período do dia”.

#### 5.2.2. Dimensão Espetador

A dimensão Espetador foi criada com o recurso à fonte de dados “espetadores.csv”. Esta fonte continha 8 campos: “ID”, “Código”, “Género”, “Região”, “Classe Social”, “Escala Etária”, “Dona de Casa” e “Data”.

Destes campos, 6 foram reaproveitados para a criação da dimensão: “Código”, “Género”, “Região”, “Classe Social”, “Escala Etária” e “Dona de Casa”.

#### 5.2.3. Dimensão Programa

Relativamente à dimensão Programa, os dados que a constituem são provenientes de duas fontes iniciais:

- os ficheiros de formato “.pet”, de onde são retiradas todas as principais informações sobre cada programa visualizado;
- o ficheiro “tipologia.tsv”, que serviu para fazer a correspondência descritiva com os tipos, categorias e géneros de cada programa.

Assim, a dimensão Programa possui 6 campos, para além do campo referente à sua chave substituta: “nome geral”, “nome específico”, “canal”, “tipo”, “categoria” e “género”.

#### 5.2.4. Tabela de Factos Audiências

A tabela de factos possui cinco atributos, sendo que quatro são chaves estrangeiras para as dimensões criadas (“Espetador”, “Programa”, “Data Início” e “Hora Início”), e o último atributo é a medida numérica definida, a duração.

### 5.3. Carregamento do Data Warehouse

Esta etapa consiste em estruturar e carregar os dados modificados para uma base de dados relacional. O carregamento pode ser simples (reescrever dados novos por cima de antigos) ou mais completo em termos de dados históricos (mantendo um registo de todas as alterações efetuadas).

Para este efeito, foi utilizado o *SQL Server Management Studio* (SSMS), mais concretamente a ferramenta *Integration Services*, seguindo o modelo dimensional. Posteriormente, utilizou-se a ferramenta *Analysis Services* do *SQL Server Business Intelligence* para o carregamento e construção do cubo de dados.

Esta ferramenta tem como principais vantagens admitir conjuntos de dados de grande dimensão, assim como permitir um controlo preciso sobre as dimensões e medidas, incluindo a definição de hierarquias de atributos.

Os programas utilizados nesta etapa e respetivas responsabilidades são sumarizados na tabela 21.

Tabela 21 - Responsabilidades da etapa "Carregamento" do sistema ETL e de cada programa

Programa	Responsabilidade	Entrada	Saída
<i>Microsoft SQL Server Management Studio 2008</i>	→ <b>Distribuição</b> - Criação das Tabelas em Linguagem SQL	Comandos SQL	Tabelas de factos e dimensões
<i>Import and Export Data (32-bits)</i>	→ <b>Distribuição</b> - Importação dos dados		
<i>SQL Server Business Intelligence Development Studio (Microsoft Visual Studio 2008)</i>	→ <b>Distribuição</b> - Importação dos dados → <b>Entrega</b> - Transferência de dados para o cubo de dados		

#### 5.3.1. SQL Server Management Studio

Antes da realização de qualquer atividade de carregamento de dados, é indispensável a conexão à base de dados com as credencias fornecidas pelo docente.

Depois de criadas as tabelas de factos e dimensões, o passo seguinte passa por criar as tabelas correspondentes na base de dados relacional, usando a linguagem SQL, para depois ser efetuado o carregamento dos dados.

Nas várias capturas de ecrã que se seguem são demonstrados os códigos SQL utilizados para a criação das diferentes tabelas.

```
CREATE TABLE dimHorario (
  id NUMERIC(6,0),
  [horario completo] NVARCHAR(MAX) NOT NULL,
  [periodo do dia] NVARCHAR(MAX) NOT NULL,
  hora NUMERIC(2,0) NOT NULL,
  minutos NUMERIC(2,0) NOT NULL,
  segundos NUMERIC(2,0) NOT NULL,
  constraint pk_dimHorario
    primary key (id),
  constraint ck_dimHorario
    check (id>=0)
);
```

Captura de Ecrã 1 - Comando SQL para criar a tabela referente à dimensão Horário

```
CREATE TABLE dimPrograma (
  id NUMERIC(9,0),
  [nome geral] NVARCHAR(MAX) NOT NULL,
  [nome específico] NVARCHAR(MAX) NOT NULL,
  canal NUMERIC(2,0) NOT NULL,
  tipo NVARCHAR(MAX) NOT NULL,
  categoria NVARCHAR(MAX) NOT NULL,
  genero NVARCHAR(MAX) NOT NULL,
  constraint pk_dimPrograma
    primary key (id),
  constraint ck_dimPrograma_id
    check (id>0)
);
```

Captura de Ecrã 2 - Comando SQL para criar a tabela referente à dimensão Programa

```
CREATE TABLE dimEspetador (
  id NUMERIC(9,0),
  [chave supernatural espetador] NUMERIC (9,0),
  codigo NUMERIC(9,0) NOT NULL,
  genero NVARCHAR(MAX) NOT NULL,
  [escala etario] NVARCHAR(MAX) NOT NULL,
  regio NVARCHAR(MAX) NOT NULL,
  estatuto1 NVARCHAR(MAX) NOT NULL,
  ocupacao1 NVARCHAR(MAX) NOT NULL,
  estatuto2 NVARCHAR(MAX) NOT NULL,
  ocupacao2 NVARCHAR(MAX) NOT NULL,
  [dona de casa] NVARCHAR(MAX) NOT NULL,
  [data inicio] DATE,
  [data fim] DATE,
  [em vigor] NVARCHAR(MAX),
  constraint pk_dimEspetador
    primary key (id),
  constraint ck_dimEspetador_id
    check (id>0)
);
```

Captura de Ecrã 3 - Comando SQL para criar a tabela referente à dimensão Espetador

```
CREATE TABLE dimData (
  id NUMERIC(8,0),
  [data completa] DATE NOT NULL,
  [dia do mes] NUMERIC(2,0) NOT NULL,
  mes NUMERIC(2,0) NOT NULL,
  [nome do mes] NVARCHAR(MAX) NOT NULL,
  ano NUMERIC(4,0) NOT NULL,
  [dia da semana] NVARCHAR(MAX) NOT NULL,
  [semana do ano] NUMERIC(2,0) NOT NULL,
  [fim de semana] NVARCHAR(MAX) NOT NULL,
  [indicador feriado] NVARCHAR(MAX) NOT NULL,
  [indicador data comemorativa] NVARCHAR(MAX) NOT NULL,
  [nome data comemorativa] NVARCHAR(MAX) NOT NULL,
  constraint pk_dimData
    primary key (id),
  constraint ck_dimData
    check (id>0)
);
```

Captura de Ecrã 4 - Comando SQL para criar a tabela referente à dimensão Data

```
CREATE TABLE factAudiencias (
  espetador NUMERIC(9,0),
  programa NUMERIC(9,0),
  [data inicio] NUMERIC(8,0),
  [hora inicio] NUMERIC(6,0),
  duracao NUMERIC(10,0) CONSTRAINT nn_factVenda_duracao NOT NULL,

  CONSTRAINT pk_factAudiencias
    PRIMARY KEY (espetador, programa, [data inicio], [hora inicio]),

  CONSTRAINT fk_factAudiencias_espetador
    FOREIGN KEY (espetador)
    REFERENCES dimEspetador(id),

  CONSTRAINT fk_factAudiencias_programa
    FOREIGN KEY (programa)
    REFERENCES dimPrograma(id),

  CONSTRAINT fk_factAudiencias_data
    FOREIGN KEY ([data inicio])
    REFERENCES dimData(id),

  CONSTRAINT fk_factAudiencias_horario
    FOREIGN KEY ([hora inicio])
    REFERENCES dimHorario(id)
);
```

Captura de Ecrã 5 - Comando SQL para criar a tabela de factos

Depois de criadas as tabelas em SQL, pode ser feito o carregamento de dados para a base de dados. O processo de carregamento dos dados foi realizado no *SQL Server Business Intelligence Development Studio*.

### 5.3.2. SQL Server Business Intelligence Development Studio

O processo de extração, transformação, e carregamento de dados para a base de dados relacional pode ser realizado com mais precisão e controlo através da criação de um *Integration Services Project*.

#### 5.3.2.1. Integration Services

Recorrendo a esta ferramenta, começámos por criar um *Integration Services Project* denominado por “AudenciasTelevisivas”.

Uma vez criado, foram seguidos uma série de passos para configurar corretamente as propriedades do projeto:

1. Criação de uma *OLE DB Connection* (“Destino de dados SQL Server”) correspondente ao destino dos dados;

ConnectionManagerType	OleDb
ConnectionString	Data Source=CACILHEIRO\DIFFUL;User ID=IPAI07
DataSourceID	
DelayValidation	False
Description	
Expressions	
ID	{23E92B69-1DBA-4B2C-AB50-E6999ABA0AFC}
InitialCatalog	IPAI07BD
Name	Destino de dados SQL Server
Password	*****
RetainSameConnection	False
ServerName	CACILHEIRO\DIFFUL
SupportsDTCTransactions	True
UserName	IPAI07

Captura de Ecrã 1 - OLE DB Connection correspondente ao destino dos dados.

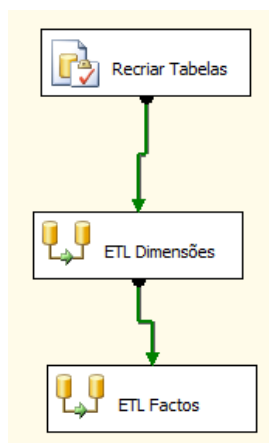
2. Criação de uma *File Connection* correspondente à fonte dos dados;
3. *Execute SQL Task*;

<b>General</b>	
Name	Recriar Tabelas
Description	Execute SQL Task
<b>Options</b>	
TimeOut	0
CodePage	1252
<b>Result Set</b>	
ResultSet	None
<b>SQL Statement</b>	
ConnectionType	OLE DB
Connection	Destino de dados SQL Server
SQLSourceType	File connection
FileConnection	TabelasFacto_e_Dimensao.sql
IsQueryStoredProcedure	False
BypassPrepare	True

Captura de Ecrã 2 - Detalhes da caixa referente à opção "Execute SQL Task"



4. *Data Flow Task* para as Dimensões e Tabela de Factos (só é possível transferir dados para as tabelas de factos depois de transferir para as tabelas de dimensões);



Captura de Ecrã 3 – Data Flow

Finalizados estes passos, as tabelas criadas no SQL Server foram carregadas com os dados provenientes dos ficheiros “.tsv”: “dimHorario.tsv”, “dimEspetador.tsv”, “dimPrograma.tsv”, “dimData.tsv” e “factosAudiencias.tsv”.

Com os dados carregados, foi possível proceder à construção do cubo de dados.

#### 5.3.2.2. Analysis Services

Antes de executar os passos relativos à construção do cubo de dados propriamente dito, é necessário **criar um novo projeto** e **definir as permissões de uso** do cubo. Este novo projeto, de tipo *Analysis Services*, deve ser criado no âmbito da solução que contém o projeto de *Integration Services*.

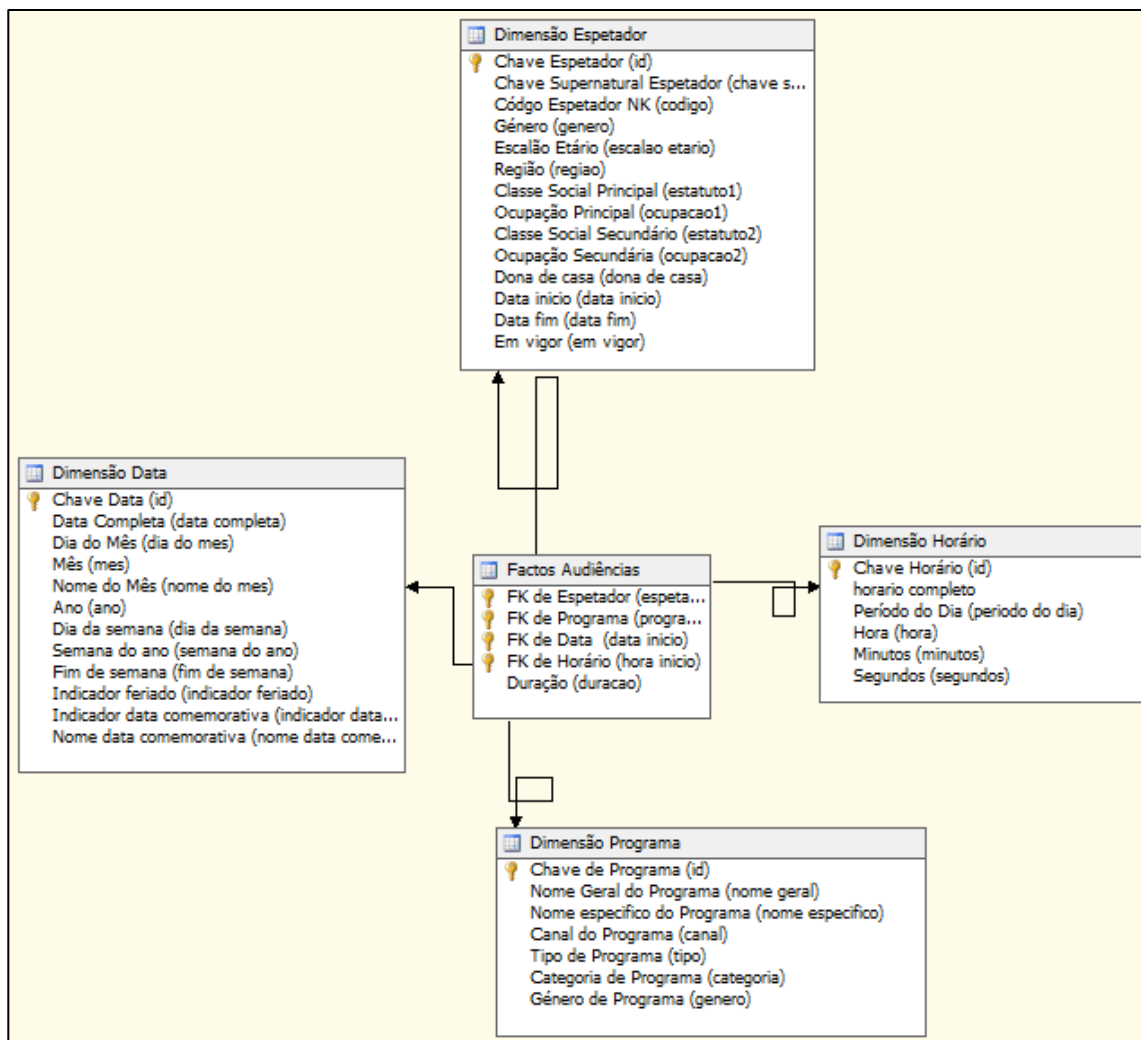
Uma vez criado o novo projeto de tipo *Analysis Services*, de seu nome “Cubo AudienciasTelevisivas IPAI07” e definidas as permissões de segurança, começámos por **identificar as fontes de dados** que irão sustentar o cubo de dados, fontes essas que irão fornecer os dados atómicos sobre medidas e dimensões.

Neste caso, existe apenas uma fonte, que é a base de dados SQL Server criada com os comandos descritos em 5.3.1. A esta fonte de dados foi dado o nome “FonteDados IPAI07”, e o servidor é “CACILHEIRO\DIFCUL”.

Definida a fonte de dados e configurado o repositório do cubo de dados, o próximo passo consiste em **selecionar as tabelas que fornecem os dados** às tabelas de dimensões e de factos.

Para este efeito, foi **criada uma vista sobre as tabelas SQL**, denominada “Vista\_IPAI07”, onde são visíveis as tabelas, respetivos atributos, e as ligações formadas pelas chaves estrangeiras, sendo que o aspeto do esquema é em estrela, com a tabela de factos no centro e as de dimensões em redor.

De seguida, foram dados nomes mais inteligíveis sobre os elementos do esquema em estrela para que, mais tarde, os atributos das dimensões e as medidas da tabela de factos possam ser incluídos num relatório dinâmico com o máximo de informação de contexto possível, conforme demonstrado no esquema 2.



Esquema 2 - Esquema em estrela gerado pelo SQL Server Business Intelligence Development Studio

Posteriormente foram **definidas as dimensões** que vão permitir navegar no cubo de dados. Embora o cubo não estivesse ainda criado, nem identificadas as medidas do negócio, já nos foi possível especificar as hierarquias (ver secção 4 e subsecções) entre os atributos das dimensões.

Depois de definidas as dimensões e respetivas hierarquias, seguiu-se a etapa de **criação do cubo de dados**, para o qual também vai ser necessário identificar a tabela que guarda os factos e as respetivas medidas. Ao cubo de dados foi dado o nome **“Cubo IPAIO7”**.

Por fim, resta fazer o **deploy do cubo**, que serve para copiar os dados atómicos das tabelas de factos e dimensões em SQL para dentro do cubo, bem como para calcular e armazenar os valores agregados resultantes de todas as combinações possíveis entre atributos das dimensões.

Assim, podemos resumir o processo de criação do cubo de dados pela ferramenta *Analysis Services* nos seguintes passos:

- 1) criação do projeto e definição de permissões de segurança;
- 2) identificação da fonte de dados e configuração do repositório para o cubo de dados;
- 3) criação de vistas mais inteligíveis sobre os dados, caso necessário;
- 4) definição de dimensões e hierarquias de atributos;
- 5) criação do cubo de dados.

Depois de seguidos estes passos, será possível **compor um relatório dinâmico**. Para tal, basta arrastar medidas e atributos de dimensões para dentro relatório. Podem ainda ser aplicados diversos filtros, para apenas serem mostrados alguns valores do leque de disponíveis.

## 6. Conclusão

Este trabalho tem como foco analisar as tendências televisivas dos espectadores, sendo esse o nosso processo de negócio prioritário. Assim sendo, o intuito deste projeto é a construção de um *data warehouse*, com o objetivo de auxiliar a tomada de decisão no contexto da análise das audiências televisivas.

Foram analisados os dados disponíveis para verificar a existência de erros que devem ser tidos em conta aquando do processo de extração, transformação e carregamento para o *data warehouse*.

Depois de estabelecido que o mais adequado seria uma tabela de factos do tipo transacional, e que o grão seria “um espectador que vê um dado programa, numa dada data e hora, durante um determinado período de tempo”, passou-se à construção do *data warehouse*.

Nesta terceira fase do projeto, focámo-nos em construir (e demonstrar) o *data warehouse*, mais especificamente todo o processo de extração, transformação e carregamento de dados.

A próxima fase irá basear-se na geração de relatórios com base no cubo criado.

## 7. Bibliografia

Romeu, A. M. (2014). *A MEDIÇÃO DAS AUDIÊNCIAS TELEVISIVAS EM PORTUGAL: NOVAS PRÁTICAS, NOVOS CONSUMOS, NOVOS DESAFIOS*. Universidade Católica Portuguesa- Faculdade de Ciências Humanas

## 8. Anexos

### 8.1. Esquema de Gráficos



Esquema 3 - Conjunto de gráficos relativos a todas as fontes de dados, onde é destacado o conjunto de dados relativo aos registos do sexo feminino

