U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Modelação de um Data Warehouse e Análise de Dados Multi-dimensionais

Rui Andrade 201304902
Vanda Azevedo 201305778

Introdução

Um *data warehouse* funciona como uma base de dados orientada a solucionar problemas de decisão ou a análise de dados históricos e consolidados. Contendo dados consolidados, não é focado em operações diárias ou em transacções, como uma base de dados relacional, mas sim em torno de vários temas para permitir essa organização. A construção de um *data warehouse* pode ter como base informações de bases de dados relacionais ou de ficheiros contendo informações sobre transacções ou entradas, podendo ser realizadas operações de pré-processamento ou integração de dados. É, portanto, uma escolha ideal para a modelação dos dados que nos foram disponibilizados, de forma a agilizar a análise e revisão de dados sob uma visão mais ampla.

Descrição dos dados fornecidos

Foram-nos fornecidas tabelas contendo informações sobre serviços de táxi (**taxi_services**) e sobre praças de táxis (**taxi_stands**). A tabela relativa aos serviços de táxi continha dados para cerca de um milhão e meio de serviços, incluindo pontos de partida e fim, tempo de início e de final e qual o táxi que desempenhou tal serviço, enquanto que a tabela relativa às praças de táxi continha o nome de cada praça e a sua localização geométrica.
Adicionalmente, também utilizamos a **CAOP** - Carta Administrativa Oficial de Portugal, disponibilizada pela Direcção Geral do Território e que contém informações sobre a divisão administrativa do território português. Antes de utilizarmos a CAOP, foi necessário converter o modelo geométrico dos dados, originalmente 27493 (Datum 73/Modified Portuguese Grid), para o 4326 (WGS86) através do comando:

```
shp2pgsql -W "latin1" -s 27493:4326 -g geom -I caop/Cont_Freg_V5.shp public.caop | psql
```

Organização do Data Warehouse

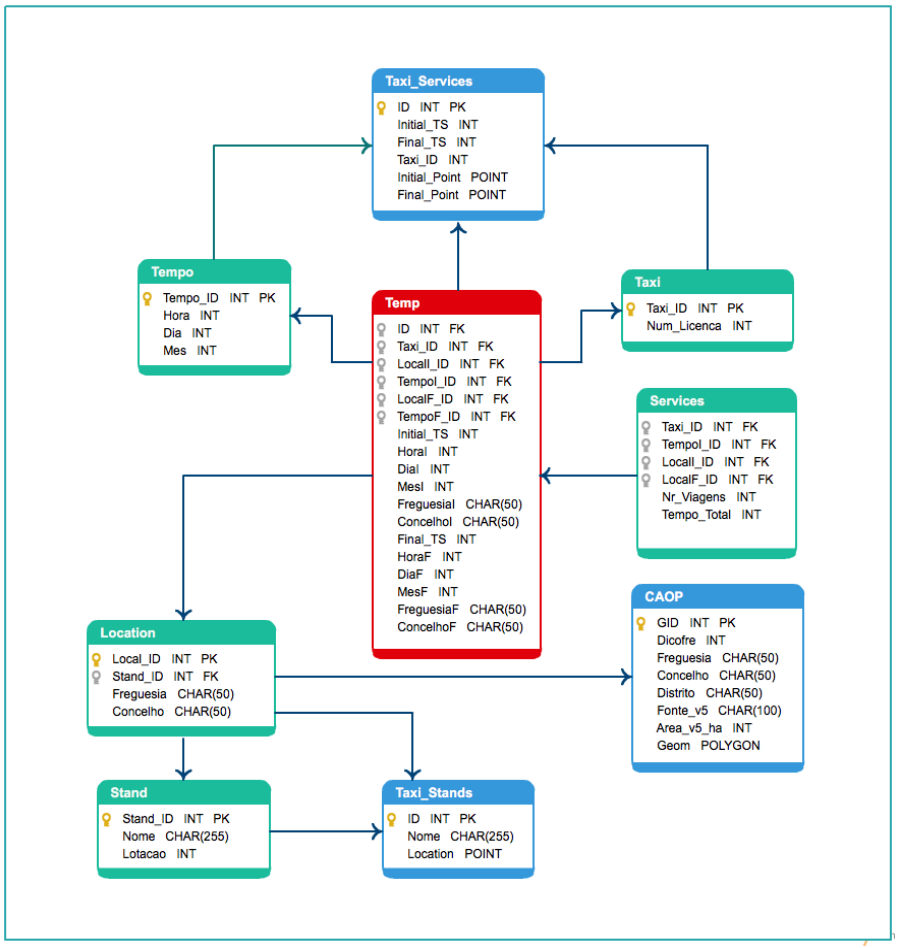
Previamente à integração dos dados no nosso *data warehouse*, procedemos à estruturação do mesmo, orientando por vários temas: tempo, táxis, localizações, praças e serviços. Cada um dos temas será representado por uma relação que conterá informação suficiente para identificar unicamente cada uma das entradas, permitindo que a tabela de serviços se relacione com as outras através de chaves externas e que agrupe entradas semelhantes.
Na relação **tempo**, a decomposição foi feita em mês, dia e hora, não sendo mais específico para permitir que ocorra algum tipo de agregação.

Na relação **taxi**, cada táxi é definido pelo seu número de licença e por um identificador gerado automaticamente.
Na relação **stand** (praça), temos o nome de cada praça e a lotação respectiva. Dado que as praças constituem um ponto, consideramos que um serviço começa ou termina numa praça caso esteja, no máximo, a 100 metros do ponto.
A relação **location** (localização) contém informação sobre freguesia, concelho e eventualmente uma ligação a uma praça noutra relação, caso exista alguma na respectiva freguesia e respectivo concelho - poderá, portanto, ter mais que uma entrada para o mesmo par de freguesia e concelho, pela possibilidade de existir mais que uma praça nessa localização, ou pela possibilidade de existirem serviços que partem de uma freguesia que contém alguma praça, mas que não partem dessa praça.
A relação **services** (serviços) liga-se com as relações de tempo, táxis e localização, permitindo que os mesmos tuplos de veículo utilizado no serviço, de tempo inicial, de localização inicial e de localização final sejam consolidados em número de viagens, evitando assim a existência de tuplos redundantes. Caso tal agrupação ocorra, temos um campo para somar o tempo total despendido neste tipo de serviços.

Modelação do Data Warehouse

O preenchimento das tabelas foi feito começando pelas que não tinham chaves externas - **tempo**, **taxi**, **stand**.
A relação **taxi** foi populada a partir dos identificadores dos táxis disponíveis na tabela **taxi_services**.
A relação **tempo** foi construída a partir dos tempos registados na tabela **taxi_services**, tendo sido feita conversão das unidades de tempo para hora, dia e mês.
A relação **stand** foi populada directamente a partir da tabela **taxi_stands**.
A relação **location** foi construída com base na tabela **taxi_stands**, na localização geométrica das mesmas, na **taxi_services** (através dos pontos de início e fim de serviço) e nos dados obtidos na **CAOP**. Com recurso à tabela **taxi_services**, foi possível saber em que freguesias e concelhos (e, eventualmente, stands) começaram ou terminaram os serviços. Ou seja, esta relação contém todas as localizações de stands e o respectivo stand_ID, e todas as localizações que, apesar de não estarem associadas a um stand (stand_ID null), são associadas a um concelho e freguesia.
A população da relação **services** foi feita com recurso a uma tabela temporária (**temp**). Com os dados da tabela **taxi_services**, procedemos à expansão das colunas (detalhando, por exemplo, que freguesia e concelho correspondem a um ponto inicial) e associação com identificadores das tabelas de dimensões. Para esse efeito, determinamos qual a chave de tempo que corresponde a cada *timestamp* e as chaves de localização para os pontos iniciais e finais. Com este nível de detalhe, podemos efectuar agregação de forma mais clara e preencher, assim, com relativa facilidade, a tabela de factos.

A maior dificuldade na construção do *data warehouse* prendeu-se com a demora na execução da *query* de construção da tabela temporária, o que acaba por não ser tão preocupante se considerarmos que um *data warehouse* é, tipicamente, construído apenas uma vez, por servir para consultas de agregação sobre dados históricos e consolidados.



Azul: tabelas fornecidas, vermelho: tabela temporária, verde: datawarehouse

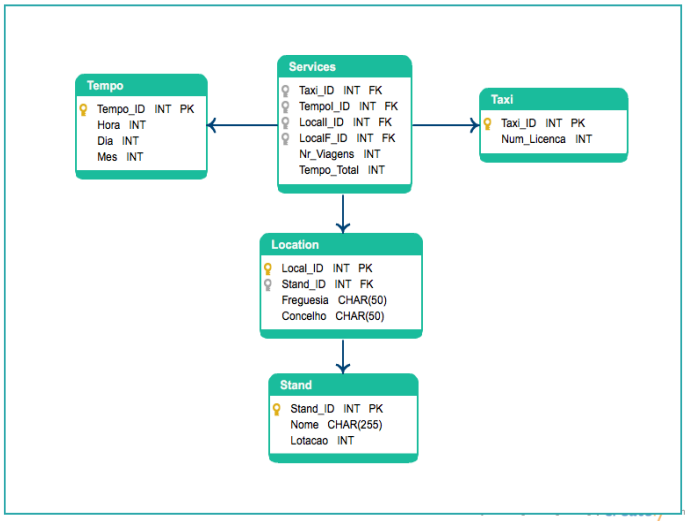


Diagrama em estrela

Data Processing

Apesar de a tabela **temp** excluir automaticamente os serviços que começam ou terminam fora de território nacional, consideramos a existência de outras situações "absurdas" presentes nos dados, tais como: viagens com durações inferiores a 3 minutos, e viagens com durações superiores a 8 horas, pelo que, estes casos foram removidos. Visto que apenas cerca de 60000 serviços começam fora do distrito do Porto, estes foram, também, removidos.
De modo a obtermos a extrairmos outro tipo de informação relativamente ao data warehouse, criamos duas funções:

- weekDay: dados um ano, um mês e um dia, indica qual o dia da semana correspondente;
- period: dada uma hora, indica ao período de tempo a que esta corresponde (madrugada]00;6], manhã]6;12], tarde]12;18], noite]18;00]).

Análise de Dados Multi-dimensionais

Dado que um data warehouse corresponde à compilação de dados de diversas bases de dados diferentes, e os seus dados são manipulados de forma conveniente, a análise de dados com recurso a um data warehouse encontra-se facilitada, pois os dados já se encontram disponíveis para acesso e não são voláteis. Os dados obtidos de um data warehouse poderão ser usados em tomadas de decisões.

Como primeiro passo na análise de dados, decidimos averiguar quais as praças de táxis mais concorridas, por ser uma medida indicativa dos locais onde há maior procura por um táxi e para tentar perceber qual o motivo por trás de uma maior afluência a certas praças. Optámos por filtrar pelas 10 praças mais concorridas para fazer uma análise mais rigorosa e relevante dos dados.

```
SELECT St.Nome, SUM(Se.Nr_Viagens)
FROM Services Se, Location L, Stand St
    WHERE Se.LocalI_ID = L.Local_ID
      AND L.Stand_ID = St.Stand_ID
            GROUP BY 1
     ORDER BY 2 DESC LIMIT 10;
```

Praça	Total de serviços
Campanhã	35.888
São Bento	23.047
Ribeira	20.149

Clérigos	17.558
Batalha	15.491
Hospital São João	13.894
Carregal	12.230
Lordelo	11.851
Bom Sucesso	10.559
Brasília	10.243

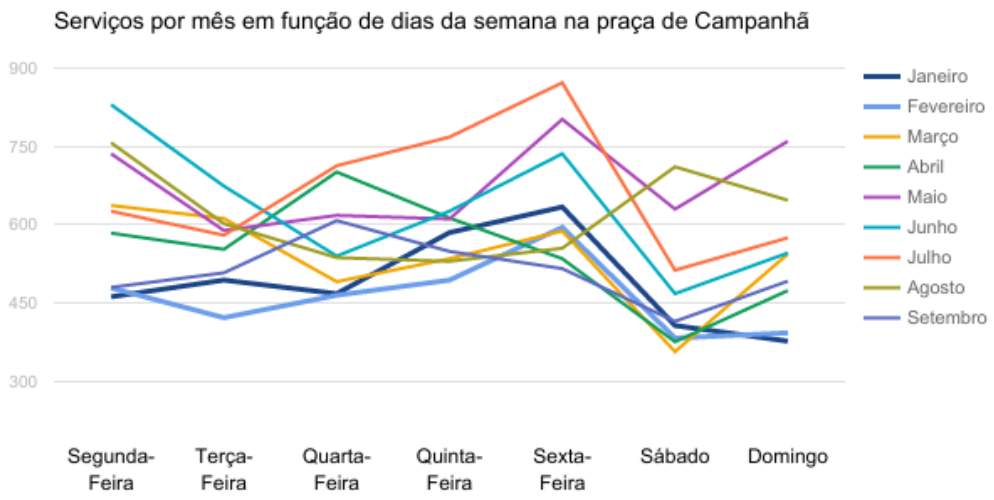
Sem surpresas, as duas praças mais concorridas são a de Campanhã e a de São Bento, facto facilmente explicado pela proximidade com as duas maiores estações ferroviárias do distrito do Porto e com estações de metro. A estação de Campanhã está, ainda, localizada perto de várias paragens de autocarro e é uma das maiores praças de táxis do distrito, estando, portanto, sempre pronta a servir os clientes. As estações da Ribeira, dos Clérigos e da Batalha localizam-se em áreas preferidas pelos turistas, não sendo também surpreendente a afluência registada às mesmas.

Como passo seguinte, e também para exemplificar que tipo de análises podem ser feitas de forma individual a cada praça, decidimos tentar perceber que factores podem influenciar a afluência à praça de Campanhã, por ser a mais concorrida de todas. Inicialmente, verificamos a distribuição dos serviços por mês e dia. Dada a extensão da tabela gerada, optámos por resumir a apresentação dos dados no relatório em distribuição de serviços por mês e apresentar a distribuição de serviços ao longo de um mês em função dos dias de semana.

```
SELECT T.Mes, T.Dia, SUM(Se.Nr_Viagens)
FROM Services Se, Location L, Stand St, Tempo T
        WHERE Se.TempoI_ID = T.Tempo_ID
         AND Se.LocalI_ID = L.Local_ID
         AND L.Stand_ID = St.Stand_ID
         AND St.Nome LIKE 'Campanhã'
          GROUP BY ROLLUP (Mes, Dia)
                ORDER BY 1,2;
```

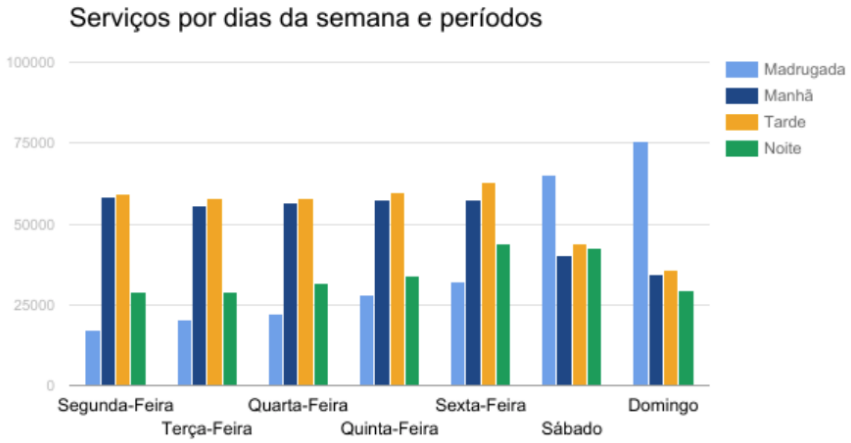
Mês	Total de serviços
Janeiro	3.427
Fevereiro	3.231
Março	3.764
Abril	3.836
Maio	4.746

Junho	4.420
Julho	4.647
Agosto	4.339
Setembro	3.478



É interessante observar que ocorre uma quebra no recurso ao serviço de táxi em Campanhã aos fins-de-semana, que poderá ser explicada pela maior afluência de alunos que vão para casa na Sexta e voltam no Domingo, e pelos "turistas de fim-de-semana". Numa visão mais geral dos dados, tentámos perceber em que dias da semana existe maior recurso ao serviço de táxi, bem como em que períodos do dia em particular. Para tal, foi óbvia a preferência por um *ROLLUP*, que apresentaria facilmente totais por dia da semana e por período de cada dia da semana.

```
SELECT weekDay(2015, Tempo.Mes, Tempo.Dia) as day_of_week, period(Tempo.Hora) as periodo,
                                SUM(Nr_Viagens) as viagens
                                   FROM Services, Tempo
                             WHERE Services.TempoI_ID = Tempo.Tempo_ID
                               GROUP BY ROLLUP(day_of_week, periodo)
          ORDER BY 1, CASE WHEN period(Tempo.Hora) LIKE 'madrugada' THEN 1
                           WHEN period(Tempo.Hora) LIKE 'manha' THEN 2
                           WHEN period(Tempo.Hora) LIKE 'tarde' THEN 3
                            WHEN period(Tempo.Hora) LIKE 'noite' THEN 4
                                        ELSE 5 END ASC;
```



É possível constatar que, ao longo da semana de trabalho, a utilização de serviços de táxi se mantém praticamente constante ao longo dos quatro períodos do dia. Contudo, ao fim-de-semana, há um pico no recurso ao táxi de madrugada, talvez explicável pelo desenvolvimento de actividades tidas como boémias e à utilização do táxi para *ridesharing* ou como opção para chegar a outros locais em segurança e com facilidade.

Com vista a perceber quais as rotas preferidas dos utilizadores do táxi e também para tentar compreender o motivo dessa preferência, obtivemos uma selecção das 10 rotas preferidas.

```
SELECT L1.Freguesia, L1.Concelho, L2.Freguesia, L2.Concelho, Sum(Se.Nr_Viagens)
                FROM Services Se, Location L1, Location L2
                    WHERE Se.LocalI_ID = L1.Local_ID
                     AND Se.LocalF_ID = L2.Local_ID
                           GROUP BY 1,2,3,4
                       ORDER BY 5 DESC LIMIT 10;
```

Freguesia de partida	**Concelho de partida**	**Freguesia de chegada**	**Concelho de chegada**	**Total de viagens**
Paranhos	Porto	Paranhos	Porto	25.115
Santo Ildefonso	Porto	Paranhos	Porto	17.856
Campanhã	Porto	Campanhã	Porto	15.931
Campanhã	Porto	Paranhos	Porto	15.521
Campanhã	Porto	Santo Ildefonso	Porto	14.424
Santo Ildefonso	Porto	Cedofeita	Porto	14.182
Paranhos	Porto	Campanhã	Porto	14.137
Santo Ildefonso	Porto	Bonfim	Porto	12.729
Paranhos	Porto	Santo Ildefonso	Porto	12.643
Santo Ildefonso	Porto	Santo Ildefonso	Porto	12.466

A forte incidência em percursos que incluam Paranhos ou Campanhã poderá ser explicada pela dimensão e população destas freguesias, sendo que são duas das maiores freguesias do Porto em termos de área e são, também, duas das freguesias com mais habitantes, sendo Paranhos a mais habitada e Campanhã a terceira mais habitada.[1]

O seguinte código tem como objectivo perceber como variam as viagens originadas nas diferentes freguesias nos diferentes períodos do dia, omitimos o concelho pois todos os pontos de origem são do concelho do Porto:

```
SELECT l1.freguesia, period(tempo.hora), SUM(services.nr_viagens) AS ViagensPeriodo
                    FROM services, location as l1, tempo
WHERE services.locali_id = l1.local_id and services.tempoi_id = tempo.tempo_id GROUP BY
                                CUBE(1,2)
        ORDER BY 2,3, CASE WHEN period(Tempo.Hora) LIKE 'madrugada' THEN 1
                    WHEN period(Tempo.Hora) LIKE 'manha' THEN 2
                    WHEN period(Tempo.Hora) LIKE 'tarde' THEN 3
                    WHEN period(Tempo.Hora) LIKE 'noite' THEN 4
                                ELSE 5 END DESC;
```

Dado que o resultado consiste numa tabela 16x5, falaremos apenas nos resultados. Estes permitem-nos concluir que a maioria dos serviços origina de Vitória durante a madrugada, talvez pela proximidade desta a bares e cafés. A freguesia de Santo Ildefonso é o ponto de origem que mais faz serviços, talvez por integrar o centro histórico. E, como visto anteriormente, os períodos de mais actividade são a manhã e a tarde.

Com recurso ao *CUBE*, também é possível fazermos o mesmo tipo de análise relativamente aos percursos para cada período do dia:

```
SELECT l1.freguesia, l1.concelho, l2.freguesia, l2.concelho, period(tempo.hora),
                    SUM(services.nr_viagens) AS ViagensPeriodo
                from services, location l1, location l2, tempo
                    where services.locali_id = l1.local_id
                    and services.tempoi_id = tempo.tempo_id
                     and services.localf_id = l2.local_id
                          group by cube(1,2,3,4,5);
```

[1] Martins, I., Ferreira, C., Rocha, E. e Gomes, M. (2014). *Censos 2011 - Mudanças Demográficas*. [PDF] Porto: Câmara Municipal do Porto, p.6. Disponível em: http://www.cm-porto.pt/assets/misc/img/PDM/Revisao_PDM/Estudos_base/Censos2011_Mudancas_demograficas_2014.pdf [Acedido a 21 de Maio de 2017].