Association for Information Systems
AIS Electronic Library (AISeL)

2018 Proceedings | Portugal (CAPSI)

2018

Data warehouse for the monitoring and analysis of water supply and consumption

José Soares
Instituto Politécnico do Cávado e do Ave, a9639@alunos.ipca.pt

Patrícia Leite
Instituto Politécnico do Cávado e do Ave, patricialeite@ipca.pt

Paulo Teixeira
Instituto Politécnico do Cávado e do Ave, pteixeira@ipca.pt

Nuno Lopes
Instituto Politécnico do Cávado e do Ave, nlopes@ipca.pt

Joaquim P. Silva
Instituto Politécnico do Cávado e do Ave, jpsilva@ipca.pt

Follow this and additional works at: https://aisel.aisnet.org/capsi2018

Recommended Citation
Soares, José; Leite, Patrícia; Teixeira, Paulo; Lopes, Nuno; and Silva, Joaquim P., "Data warehouse for the monitoring and analysis of water supply and consumption" (2018). *2018 Proceedings*. 3.
https://aisel.aisnet.org/capsi2018/3

This material is brought to you by the Portugal (CAPSI) at AIS Electronic Library (AISeL). It has been accepted for inclusion in 2018 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Armazém de dados para monitorização e análise de distribuição e consumo de água

Data warehouse for the monitoring and analysis of water supply and consumption

José Soares, Instituto Politécnico do Cávado e do Ave, Portugal, a9639@alunos.ipca.pt
Patrícia Leite, Instituto Politécnico do Cávado e do Ave, Portugal, patricialeite@ipca.pt
Paulo Teixeira, Instituto Politécnico do Cávado e do Ave, Portugal, pteixeira@ipca.pt
Nuno Lopes, Instituto Politécnico do Cávado e do Ave, Portugal, nlopes@ipca.pt
Joaquim P. Silva, Instituto Politécnico do Cávado e do Ave, Portugal, jpsilva@ipca.pt

Resumo

A água é um recurso essencial cada vez mais escasso. As atuais redes de distribuição de água potável estão a ser sujeitas a uma pressão crescente devido ao aumento constante do consumo e continuam a ter níveis de perdas de água elevados. Com o objetivo de reduzir as perdas de água e melhorar a gestão do consumo de água, a EAmb - Esposende Ambiente, E.M. está a implementar um armazém de dados de fornecimento e consumo de água. Os dados disponíveis irão permitir calcular indicadores de desempenho para monitorizar e analisar o consumo e distribuição de água no concelho de Esposende.

Palavras-chave: rede de distribuição de água; armazém de dados; processo ETL

Abstract

Water is an essential resource that is increasingly scarce. Existing water supply networks are highly stressed due the increasing water consumption and the high quantity of water losses. In order to reduce water losses and improve water consumption management, EAmb - Esposende Ambiente, E.M. is implementing a data warehouse for storing water supply and consumption data. The available data will be used to monitor and analyze water supply and consumption in Esposende county.

Keywords: *water supply system, data warehouse; ETL process*

1. INTRODUÇÃO

A água é essencial à existência de vida na Terra e o acesso a água potável é um direito legal que todos os seres humanos deveriam usufruir. No entanto, atualmente cerca de 2.1 biliões de pessoas carecem de serviços de distribuição de água potável (United Nations, s.d.).

A EAmb - Esposende Ambiente, E.M., entidade na qual decorre o desenvolvimento deste projeto, é responsável pelos sistemas públicos de captação distribuição de água no município de Esposende. Este serviço traduz-se na aquisição de água junto da entidade Águas do Norte, com posterior

ISSN 2183-489X

distribuição pelas habitações dos munícipes e serviços industriais. Entre o volume adquirido e o faturado aos utilizadores, encontra-se um desfasamento que ronda os 25%, valor este abaixo do nível nacional fixado em 29.8% (ERSAR, 2017, p. 19). Ainda assim, isto significa que a cada 100 litros, cerca de 25 litros são considerados como perdas, potencialmente causadas por fugas, transgressões de abastecimento, roturas, etc. Estes fatores são vulgarmente identificados e classificados como perdas reais e perdas aparentes.

As perdas reais refletem a eficiência do sistema de distribuição e estão presentes em todos os sistemas, sendo praticamente impossíveis de erradicar por completo. Estas perdas ocorrem geralmente devido à utilização de materiais fracos, instalações descuidadas, excesso de pressão, corrosão dos materiais, falta de manutenção, etc. As perdas aparentes, como a própria descrição indica, não resultam de fugas físicas de água, mas sim de medições incorretas ou utilização não autorizada. Regularmente, a medição incorreta pode resultar de instalações incorretas, tamanhos inadequados, desgaste com o tempo, etc. A informação sobre perdas aparentes é crucial para o distribuidor e chegam a atingir um nível de importância elevado já que representam um efeito negativo superior (Thornton, Sturm, e Kunkel, 2008, p. 5).

Numa visão de perspetiva futura e melhoria do serviço de distribuição de água, a EAmb - Esposende Ambiente, E.M. pretende desenvolver um sistema analítico que, integrando os dados de que dispõe em vários sistemas, permita obter um conhecimento mais concreto dos fatores que influenciam ou possam influenciar a eficiência deste serviço. Este artigo descreve o primeiro passo no desenvolvimento do sistema, que consiste na implementação de um armazém de dados ou *data warehouse* que suporte a obtenção de indicadores de desempenho do serviço e a extrações de padrões de funcionamento do mesmo. Como tal, é fundamental a implementação de um processo de ETL (Extract-Transform-Load), capaz de conjugar distintas fontes de dados por forma a preencher o armazém de dados com dados de valor acrescentado.

Na secção seguinte, são apresentados diversos estudos, relevantes no contexto deste projeto, sobre a monitorização e detenção de perdas de água nos sistemas de distribuição de água de consumo. A terceira e quarta secções descrevem os requisitos de informação e a modelação do armazém de dados. Na quinta secção, é apresentada a implementação do processo de ETL (*Extract-Transform-Load*) e os constrangimentos que foram encontrados. Por último, apresentam-se as conclusões desta etapa do projeto e o trabalho futuro.

2. TRABALHO RELACIONADO

Apesar do impacto económico e do aumento das dificuldades de gestão dos sistemas de distribuição de água de consumo (González-Gómez, García-Rubio, e Guardiola, 2011), não têm sido tomadas fortes medidas para a redução das perdas de água (Kanakoudis, Tsitsifli, Samaras, e Zouboulis,

2013). Este pode ser o motivo pelo qual é tão difícil encontrar trabalhos sobre a implementação de armazéns de dados e utilização de indicadores de desempenho na área dos sistemas de distribuição de água de consumo. Nos vários momentos de pesquisa que foram realizados ao longo deste trabalho, não foi encontrado qualquer trabalho de implementação de armazéns de dados no setor das águas de consumo. O desenvolvimento de armazéns de dados e sistemas analíticos não faz parte das principais prioridades para lidar com os desafios relacionados com a gestão da água de consumo (OECD, 2016).

Existem muitos estudos relacionados com a deteção de fugas, especificamente sobre a análise de valores de pressão e manutenção de redes, as chamadas perdas reais. Um estudo realizado por Puust, Zapelan, Savic e Koppel (2010) apresenta os principais métodos usados na avaliação, deteção e controlo de fugas de água. Neste estudo, encontram-se apenas algumas referências à utilização de redes neurais artificiais para realizar previsões de consumo a curto prazo (Bougadis, Adamowski, e Diduch, 2005) e a deteção de picos de caudal e fugas (S. R. Mounce, Boxall, e Machell, 2010; Stephen R. Mounce e Machell, 2006). Noutro estudo, as técnicas de mineração de dados foram usadas para determinar picos de caudal e apoiar a gestão dos ativos da infraestrutura da rede (Babovic, Drécourt, Keijzer, e Friss Hansen, 2002). Este estudo avalia a detioração e o nível de risco na rede de distribuição com base em fatores como, por exemplo, a idade, o diâmetro, o material, o terreno onde se encontra a tubagem, etc., identificando as tubagens a substituir e a necessidade de novas tubagens.

Existem muitos estudos que se centram na água não faturada. Güngör-Demirci, Lee, Keck, Guzzetta, e Yang (2018) identificam os principais fatores que influenciam as perdas de água não faturada. Vilanova, Filho e Balestieri (2014) apresentam uma revisão de literatura sobre indicadores de desempenho dos serviços de distribuição de água de consumo. No entanto, apesar de ter sido realizada uma pesquisa exaustiva, apenas foi encontrado um estudo que propõe a implementação de um armazém de dados para monitorizar a qualidade da água de consumo numa região de Taiwan, integrando informação de várias fontes (Wang e Guo, 2013).

3. REQUISITOS DE NEGÓCIO

Nesta secção é realizada uma breve descrição da estrutura e funcionamento da rede, seguida da identificação dos principais requisitos de negócio.

3.1. Rede de abastecimento de água

A rede de abastecimento de água da EAmb - Esposende Ambiente, E.M. tem sido alvo de constante atenção e melhoria através da integração de novos componentes, substituição de material deteriorado e reorganização de tubagens. A atual composição da rede destaca os reservatórios, tubagens adutoras, distribuidoras e ramais, citadas por ordem de grandeza, contadores de zona e contadores

de clientes. O relacionamento destes componentes pode ser representado em várias estruturas de árvore invertida, uma por cada reservatório. Apesar de não pertencerem EAmb - Esposende Ambiente, E.M., os reservatórios são a principal fonte de fornecimento de água. Imediatamente seguinte aos reservatórios estão presentes contadores que identificam o volume e caudal de água consumida. Posteriormente, a água é conduzida por tubagens adutoras e distribuidoras, entre as quais existem contadores de zona estrategicamente colocados. Nos pontos terminais da estruturam encontram-se os ramais e contadores de consumidor.

Para facilitar os serviços de gestão da rede, foram definidas áreas de controlo que agregam todos os consumidores afetos a determinadas tubagens de uma determinada zona geográfica, sendo representadas por um contador de zona que regista o volume de água consumido. Esta identificação permite uma maior precisão na deteção de fugas ao conseguir isolar áreas em que o contador registou um aumento brusco de consumo ou diminuição substancial de pressão. A Figura 1 representa de forma visual e sucinta esta informação.

Existem ainda outros componentes presentes na rede, com menor importância para este estudo, que ajudam a manter a estabilidade da rede: as válvulas redutoras de pressão que reduzem o fluxo de água; as sobrepressoras que aumentam a pressão em pontos mais longínquos; as descargas de fundo que são utilizadas para esvaziar tubagens; e as válvulas de corte que são acionadas em conjunto com as descargas de fundo numa eventual necessidade de reparação.

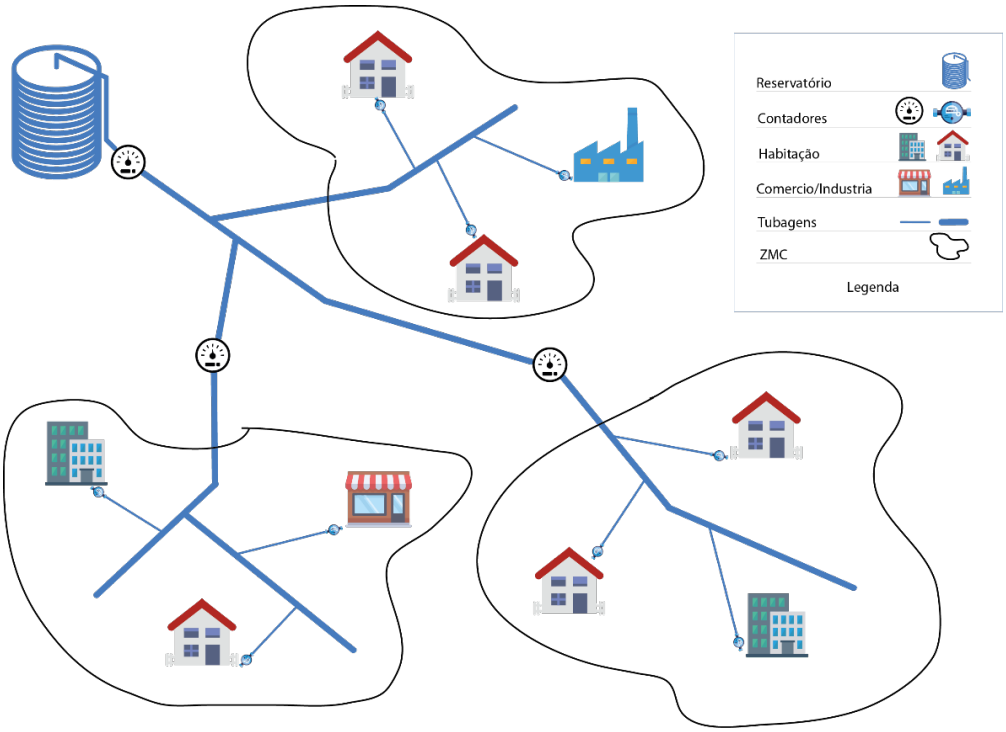

Figura 1- Esquema representativo do sistema de distribuição

3.2. Lista dos requisitos de negócio

O desenvolvimento do armazém de dados seguiu os seguintes requisitos de negócio que ilustram as necessidades de análise em termos das medidas, dimensões e tipos de análises:

- **Caudal**: o volume caudal é uma das medidas mais relevantes aquando da deteção de fugas de um sistema de distribuição de águas, apesar de ser expectável a sua variação por vários fatores, os valores tendem a ser semelhantes por certos períodos.
- **Consumo**: correspondente às medidas volume de água distribuído e consequente faturação, representa a eficiência geral do sistema de distribuição,
- **Contador**: as diversas leituras estarão sempre afetas a um contador, o qual pode corresponder a uma ou mais zonas de medição e controlo.
- **Perfil de consumidor**: a informação sobre os consumidores e o seu perfil de consumo possibilita a deteção de comportamentos fraudulentos através das variações bruscas de consumo e da identificação de valores irregulares, não compatíveis com o perfil.
- **Influência climatérica**: as variações climatéricas justificam alterações de consumo nos consumidores dependendo do tipo de atividade, especialmente em zonas agrícolas, as quais representam um segmento de consumo significativo.
- **Os períodos homólogos**: a comparação de períodos homólogos, além representar a principal fonte de indicadores de desempenho operacionais (KPI), pode servir para avaliar os investimentos efetuados na aplicação de novos materiais ou soluções.
- **Mínimos noturnos**: a análise dos valores mínimos de caudal e volume noturno pode servir como base para o ajuste de pressões, bem como deteção de fugas se relacionados com valores médios de períodos adjacentes.
- **Influência sazonal**: sendo Esposende geograficamente caracterizado por zonas costeiras e justificável afluência em época balnear, os consumos refletem em alta este acontecimento, pelo que a sua análise é de elevado interesse.

Ainda está em estudo a identificação dos perfis de utilização do armazém de dados e a validação dos indicadores chave de desempenho. Devido à necessidade de ir acumulando dados históricos no armazém de dados, decidiu-se avançar de imediato com a implementação da base de dados e do processo de ETL, tendo sido incorporados todos os dados disponíveis relativos às leituras de caudais e consumos na rede de distribuição de água.

4. MODELO DIMENSIONAL

Para preparar a modelação do armazém de dados e a especificação dos processos de ETL, foi realizado um estudo dos dados, apresentado de seguida. Esta secção inclui ainda o modelo de dados do armazém de dados implementado e arquitetura geral da solução.

4.1. Estudo de dados

Os dados utilizados para o desenvolvimento deste projeto englobam temas como caudal instantâneo, volume caudal nos reservatórios, contadores intermédios e pontos finais, áreas de afetação, condições meteorológicas e localização geográfica. As áreas de afetação dos consumidores são designadas por zonas de medição e controlo (ZMC), embora algumas delas, atualmente, ainda não tenham contador próprio.

Os dados relativos ao volume e caudal dos reservatórios e dos contadores intermédios são provenientes de bases de dados de suporte a aplicações de gestão. Os restantes dados, relativos a leituras dos contadores, são obtidos em ficheiros Excel, através de acessos *FTP* aos sistemas da entidade que fornece a água. Além de terem diversas fontes de proveniência, os dados recolhidos apresentam uma grande discrepância na frequência em que são gerados novos dados: intervalos irregulares entre 1 e 2 minutos e atualização com a mesma periocidade para as bases de dados; intervalos regulares de 15 minutos e atualização diária em horário predefinido para os dados obtidos por *FTP*. Destas fontes destacam-se dados como volume acumulado de consumo, data e hora de medição, identificação de equipamento e, em parte dos equipamentos, caudal instantâneo.

Relativamente ao consumidor final, os dados provêm do sistema ERP da empresa, baseado na tecnologia *Informix*, e comportam informação sobre volume consumido, data de medição, tipo de consumidor, ramal de distribuição e área de afetação. Desta fonte, destaca-se a irregularidade temporal de medições de volume, com periocidade aproximada a um mês, podendo ser desfasada alguns dias, dependendo da calendarização de dias úteis ou disponibilidade do pessoal leitor.

Em complemento aos anteriormente referidos, e como ponto de ligação, os dados de caráter geográfico comportam informação sobre a área de afetação (ZMC), sobre equipamentos desde a fonte ao ramal de consumidor final, localização geográfica sobre coordenadas cardeais e ano de instalação. A fonte está assegurada por uma base de dados relacional de tecnologia Oracle, destacando-se, pela negativa, a ausência de histórico de funcionamento relativamente a alterações na rede.

Por fim, também a informação meteorológica está disponível sobre uma base de dados que tem vindo a ser preenchida diariamente em intervalos regulares de uma hora. Desta fonte, entre os diversos campos, incluem-se a temperatura, a precipitação e o selo temporal, dados que estão a ser recolhidos pela possível influência no consumo de água.

4.2. Modelo de dados

Considerando os requisitos previamente enunciados, foi concebido um modelo dimensional de dados, apresentado na Figura 2, o qual comporta diversas tabelas de dimensão e factos. No processo de ETL, foram aplicadas técnicas de tratamento das dimensões de alteração lenta para preservar o

histórico de alterações nos consumidores e nas ZMC. Apesar de diversidade de dados, o modelo comporta a informação num número reduzido de tabelas, pois várias dimensões são partilhadas entre as tabelas de factos.

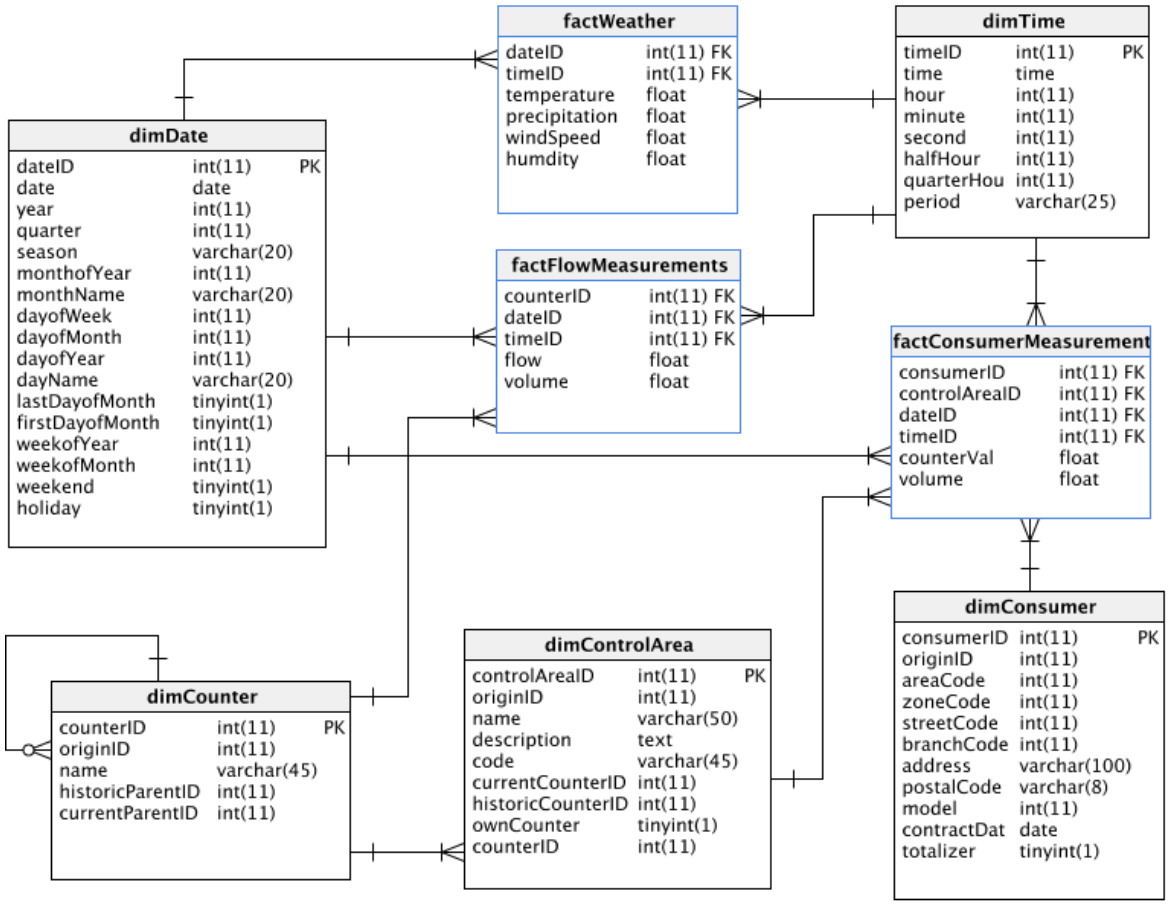

Figura 2 - Modelo de dados

O modelo exige um maior esforço no processo de ETL para potenciar os processos de análise. Identificam-se, de seguida, as tabelas dimensão excetuando as já conhecidas dimensões data e tempo:

- **DimCounter**: suporta a informação dos contadores de medição, respetiva área de afetação e posição geográfica. A particularidade desta dimensão assenta na recursividade hierárquica entre contadores, em formato de árvore.
- **DimConsumer**: o consumir revê-se representado nesta dimensão, com historial de atividade e respetivos atributos de interesse, sejam estes o tipo de cliente, localização geográfica, etc.

- **DimControlArea**: nesta dimensão encontram-se os dados relativos às ZMC inerentes aos consumidores e contadores. Esta dimensão será uma das principais fontes de ligação entre consumidores e contadores.

O modelo inclui as seguintes tabelas de factos:

- **FactFlowMeasurementes:** nesta tabela estão representados todos os factos de medições de contadores, caudal instantâneo (m^3/h) e volume (m^3), e respetiva identificação temporal. Denote-se que os intervalos temporais foram ajustados para o período mais alargado, 15 minutos concretamente, por forma a manter a uniformidade de dados.
- **FactConsumerMeasurements:** dada a exclusividade temporal das medições e ausência de valores relativos a pressão nos consumidores, nesta tabela armazena-se os factos relativos ao volume consumido expresso em m^3.
- **FactWeather:** os factos climáticos, temperatura (ºC) e precipitação (mm^3/h) estão presentes nesta tabela estando apenas diretamente relacionados com as dimensões de data e tempo.

4.3. Arquitetura da solução

A principal fonte de informação de um sistema analítico é a sua base de dados, esta deve estar devidamente preenchida com dados estruturados, percetíveis e de fácil acesso de modo a simplificar as ações de analise (Jarke, Lenzerini, Vassiliou, e Vassiliadis, 2000). Como em grande parte dos casos, os dados encontram-se dispersos por diversas fontes e formatos, foi necessário implementar todas as transformações requeridas recorrendo à ferramenta de ETL Pentaho Data Integration, conforme está representado na Figura 3.

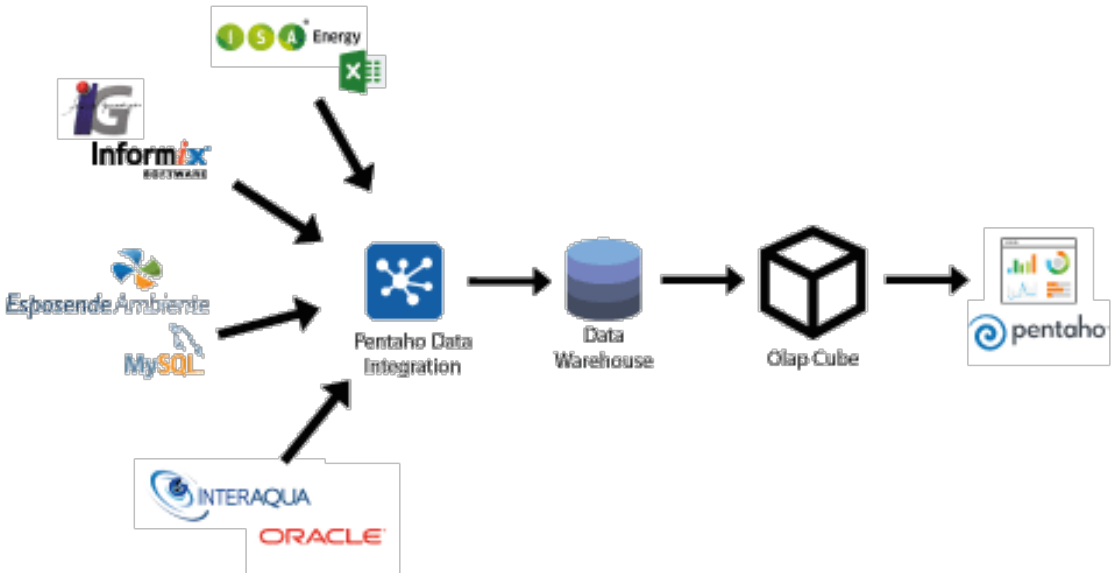

Figura 3 - Arquitetura da solução

A arquitetura comporta, além do processo de ETL, os componentes dos processos subsequentes de processamento analítico e visualização de dados. O processo ETL é responsável por este conjunto de ações, desde a extração de dados das várias fontes, passando pelo tratamento que pode englobar ações limpeza, revisão de formato, filtro, integração, etc. até atingir a fase de carregamento armazém de dados (Kimball e Caserta, 2004). A tecnologia OLAP, acrónimo de *Online Analytical Processing*, é uma abordagem de análise multidimensional que permite melhorar o desempenho das bases de dados relacionais na consulta de dados ao fornecer análises mais rápidas e com suporte a operações de *Roll-Up*, *Drill-Down* e *Pivot* (Kimball e Ross, 2013). Por fim, a servidor Pentaho BI suporta a implementação de *dashboards*, que apresentam graficamente os dados processados no cubo OLAP, e permite gerir o acesso dos utilizadores aos mesmos. A escolha do software de integração e processamento analítico foi realizada em consonância com as políticas da empresa, as quais priorizam a utilização de ferramentas *open source*.

5. IMPLEMENTAÇÃO E CONSTRANGIMENTOS

A extração de dados, apesar de diversificada por vários tipos de fontes, sejam estas FTP, Excel, Oracle, Informix e MySQL, revelou-se, de modo geral, um processo simples já que a informação se encontrava em formatos semelhantes, com exceção dos formatos dos atributos data e hora que diferem entre as fontes.

5.1. Execução do processo

Como previamente identificado, os valores dos contadores têm origens, períodos e campos diferentes, o que exige uma adaptação entre estes. A diferencia de períodos é notória, 15 minutos num caso e 1 a 2 minutos noutro. Optou-se por filtrar os intervalos mais diminutos por forma a obter o resultado mais aproximado possível do intervalo superior. O volume de consumo é fornecido num valor de natureza acumulativa, sendo necessário transformar este valor numa relação de consumo por intervalo de tempo, o que exigiu uma transformação. Apesar da ausência de caudal instantâneo em alguns equipamentos, este valor pode ser calculado pela relação entre intervalos de tempo e volume recorrendo à seguinte formula $\frac{Vf-Vi}{\Delta T}$. A identificação da sua ZMC foi efetuada através de identificador próprio da entidade fornecedora do serviço, o qual está indicado na fonte de dados geográfica.

Os dados dos consumidores provêm todos da mesma fonte de dados e, tal como os contadores, o volume é fornecido como um valor acumulado, pelo que é necessário aplicar o mesmo método, referido no parágrafo anterior, para obter o valor de consumo relativo ao intervalo. Neste caso, são ignorados os valores de caudal instantâneo, uma vez que o intervalo de medições é alargado e não justifica o seu cálculo.

Um ponto importante aquando do carregamento de valores de consumo consta na existência de consumidores "pai", mais concretamente em casos como os prédios de habitação e condomínios, em que, além do contador de cada consumidor, está instalado previamente contador geral. A medição deste inclui, além do consumo de área comum, todo o consumo gerado pelos moradores. Por questões de análise, optou-se por subtrair ao valor geral a soma dos consumidores do prédio de forma a viabilizar operações acumulativas e evitar dupla contagem.

Ainda no tratamento dos dados de consumidor, aquando da integração dos dados geográficos com as ZMC, foi detetada uma discrepância entre as ZMC afetas ao consumidor no sistema ERP e a ZMC afeta ao ramal do consumidor na base de dados geográficos. Para eliminar esta discrepância decidiu-se extrair uma listagem destas incoerências a cada execução do processo de ETL e, posteriormente, enviar esta ao pessoal responsável para identificação e correção do problema.

Relativamente à fonte de dados geográfica surgiu a necessidade de transformação de valores geográficos do formato *Datum 73*, presente na fonte de dados, para o formato global *WGS84*. Esta operação foi executada com recurso à biblioteca PROJ.4. No que diz respeito aos dados meteorológicos a transformação foi nula, sendo apenas necessário relacionar as dimensões data e tempo e filtrar os campos de interesse.

A periodicidade de execução diária dos processos de ETL já implementados não é homogénea, tanto na periodicidade como na janela temporal de execução, dependendo de cada uma das fontes de dados, como se pode observar na Tabela 1.

Origens de dados	**Início**	**Fim**	**Intervalo**
DimConsumer	08:30 h	16:30:h	01:00 h
DimCounter, DimControlArea	08:30 h	18:30 h	01:00 h-
FactConsumerMeasurements…	16:30 h	16:30 h	-
FactWeather	-	-	01:00 h
FactFlowMeasurements	-	-	00:15 h

Tabela 1- Periocidade de execução

A dimensão "DimConsumer" é atualizada em intervalos regulares durante o período de funcionamento da empresa para o domínio público, de forma a registar todas as alterações e a criação de novos registos. No mesmo intervalo de execução, a dimensão "DimCounter" e dimensão "DimControloArea" são atualizadas em função do horário do pessoal técnico.

A tabela de factos "FactConsumerMeasusements" é apenas atualizada uma vez ao dia, após descarga por parte do pessoal leitor. A tabela de factos "FactFlowMeasurements" é atualizada continuamente sem interrupções com as leituras dos contadores, em intervalos de 15 minutos de forma a garantir uma baixa latência dos dados, embora parte destes registos apenas surjam em horas determinadas pela entidade gestora da fonte de dados. Para terminar, a tabela "FactWeather" é atualizada em

intervalos de uma hora, coincidente com a hora de leitura da fonte de dados das leituras dos contadores, reduzindo assim o esforço de execução e a necessidade de armazenamento de registos.

Na primeira execução do processo de ETL foram carregados para as tabelas de factos do armazém de dados cerca de 1,5 milhões de registos na tabela "FactConsumerMeasurements", 10 mil registos para a tabela "FactWeather" e 3 milhões de registos na tabela "FactFlowMeasurements".

5.2. Constrangimentos

No decorrer da implementação foi mantida uma análise e validação constante dos valores produzidos pela comparação dos indicadores manuais existentes. Esta abordagem permitiu detetar alguns erros em valores relativos a contadores afetos às ZMC. O problema consistia na descrição aplicada às ZMC que detinham parte do nome "Restante". Afinal, o contador identificado nessa ZMC não existia na realidade, sendo referente ao nó correspondente a essa área geográfica. Isto significa que para obter o consumo dessa área é necessário subtrair a soma das restantes áreas ao valor total do contador que abrange todas as áreas. No seguimento da deteção deste erro, foram identificadas várias ZMC em que não será possível precisar o valor de consumo, uma vez que o mesmo contador inclui várias ZMC sem contador próprio.

Apesar do processo de integração ter corrido bem, verificou-se que o histórico de leituras dos contadores é muito reduzido. Enquanto que do lado dos consumidores estão disponíveis dados desde o inicio do ano de 2010, nos contadores, os primeiros registos encontram-se dispersos por datas entre os meses de junho e setembro de 2017. Esta limitação poderá vir a ser contornada se a entidade que gere o fornecimento da água for capaz de disponibilizar dados relativos a períodos anteriores.

6. CONCLUSÕES E TRABALHO FUTURO

Tendo já sido concluída a implementação do processo de ETL de alimentação do armazém de dados, apesar dos problemas detetados atrás descritos, o objetivo foi plenamente atingido. Neste momento, o armazém de dados está a ser continuamente alimentado com dados de qualidade que irão suportar a realização da próxima etapa deste projeto.

Os dados carregados no armazém de dados são rigorosos e estão devidamente estruturados. A única situação que irá dificultar a análise dos dados prende-se com o facto de alguns contadores integrarem várias ZMC sem contador próprio. Este problema irá ser certamente resolvido no futuro. Para mitigar este problema, poderá ser realizada uma estimativa do consumo dessas ZMC usando os valores históricos de consumos dos utilizadores para determinar uma percentagem de afetação do valor restante a cada ZMC.

A próxima etapa deste projeto irá consistir na implementação dos processos analíticos que permitam calcular indicadores e explorar e visualizar graficamente os dados recolhidos, recorrendo à

tecnologia OLAP e ao desenvolvimento de *dashboard* em ferramentas de *business analytics*. Uma terceira etapa deste projeto consistirá na realização de processos de mineração de dados e implementação de processos de aprendizagem automática para deteção de perdas de águas, previsão de consumos, deteção de irregularidades, previsão de ruturas no fornecimento, etc.

REFERÊNCIAS

Babovic, V., Drécourt, J.-P., Keijzer, M., e Friss Hansen, P. (2002). A data mining approach to modelling of water supply assets. *Urban Water*, *4*(4), 401–414. http://doi.org/10.1016/S1462-0758(02)00034-1

Bougadis, J., Adamowski, K., e Diduch, R. (2005). Short-term municipal water demand forecasting. *Hydrological Processes*, *19*(1), 137–148. http://doi.org/10.1002/hyp.5763

ERSAR. (2017). *Caraterização do setor de águas e resíduos (volume 1)*. *Relatório Anual dos Serviços de Águas e Resíduos em Portugal*. Obtido em 2017/05/20 de http://www.ersar.pt/pt/publicacoes/relatorio-anual-do-setor

González-Gómez, F., García-Rubio, M. A., e Guardiola, J. (2011). Why is non-revenue water so high in so many cities? *International Journal of Water Resources Development*, *27*(2), 345–360. http://doi.org/10.1080/07900627.2010.548317

Güngör-Demirci, G., Lee, J., Keck, J., Guzzetta, R., e Yang, P. (2018). Determinants of non-revenue water for a water utility in California. *Journal of Water Supply: Research and Technology - Aqua*, *67*(3), 270–278. http://doi.org/10.2166/aqua.2018.152

Jarke, M., Lenzerini, M., Vassiliou, Y., e Vassiliadis, P. (2000). *Fundamentals of Data Warehouses*. Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-662-04138-3

Kanakoudis, V., Tsitsifli, S., Samaras, P., e Zouboulis, A. (2013). Assessing the performance of urban water networks across the EU Mediterranean area: The paradox of high NRW levels and absence of respective reduction measures. *Water Science and Technology: Water Supply*, *13*(4), 939–950. http://doi.org/10.2166/ws.2013.044

Kimball, R., Caserta, J., Kimball, R., & Caserta, J. (2004). The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning Conforming, and Delivering Data. Wiley. Wiley. http://doi.org/10.1017/CBO9781107415324.004

Kimball, R., e Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). Wiley Publishing.

Mounce, S. R., Boxall, J. B., e Machell, J. (2010). Development and Verification of an Online Artificial Intelligence System for Detection of Bursts and Other Abnormal Flows. *Journal of Water Resources Planning and Management*, *136*(3), 309–318. http://doi.org/10.1061/(ASCE)WR.1943-5452.0000030

Mounce, S. R., e Machell, J. (2006). Burst detection using hydraulic data from water distribution systems with artificial neural networks. *Urban Water Journal*, *3*(1), 21–31. http://doi.org/10.1080/15730620600578538

Nogueira Vilanova, M. R., Filho, P. M., e Perrella Balestieri, J. A. (2014, March 1). Performance measurement and indicators for water supply management: Review and international cases. *Renewable and Sustainable Energy Reviews*. Pergamon. http://doi.org/10.1016/j.rser.2014.11.043

OECD. (2016). *Water Governance in Cities*. OECD Publishing. http://doi.org/10.1787/9789264251090-en

Puust, R., Kapelan, Z., Savic, D. A., e Koppel, T. (2010). A review of methods for leakage management in pipe networks. *Urban Water Journal*, *7*(1), 25–45. http://doi.org/10.1080/15730621003610878

Thornton, J., Sturm, R., e Kunkel, G. A. (2008). *Water loss control*. McGraw-Hill.

Wang, H. C., e Guo, J. L. (2013). Constructing a water quality 2.0 OLAP system in Taiwan. *Journal of Cleaner Production*, *40*, 40–45. http://doi.org/10.1016/j.jclepro.2011.04.019