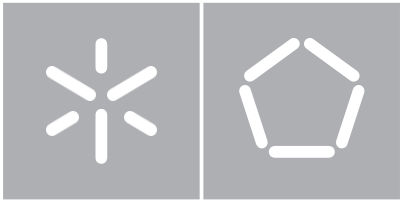
Universidade do Minho
Escola de Engenharia

Gil Gonçalves

Process Mining na Análise de Tráfego de uma Rede de Comunicações

Outubro 2018



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Gil Gonçalves

Process Mining na Análise de Tráfego de uma Rede de Comunicações

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob orientação de

Orlando Belo

Outubro 2018

Universidade do Minho

Escola de Engenharia

Departamento de Informática

Process Mining na Análise de Tráfego de uma Rede de Comunicações

Gil Gonçalves

Dissertação de Mestrado

2018

Process Mining na Análise de Tráfego de uma Rede de Comunicações

Gil Gonçalves

Dissertação apresentada à Universidade do Minho para obtenção do grau de Mestre em Engenharia Informática, elaborada sob orientação do Professor Doutor Orlando Manuel de Oliveira Belo.

2018

A minha família e amigos

Agradecimentos

Embora uma dissertação seja um trabalho individual, tal trabalho nunca seria possível sem o apoio das mais diversas pessoas que me ajudaram direta ou indiretamente na concretização desta dissertação.

Em primeiro lugar gostaria de agradecer a minha família especialmente ao meu pai e a minha mãe que tornaram possível a minha ida para a universidade. Agradeço também todo o apoio por eles demonstrado ao longo do meu percurso académico mostrando-me que por maior que seja um obstáculo é sempre ultrapassável.

A dona Alzira por me ter acolhido e aturado desde o meu primeiro ano de universidade.

Aos meus amigos de Fafe, em especial ao meu grupo "Cave Syndicate", um muito obrigado por todas as aventuras que me proporcionaram assim como o apoio demonstrado.

Aos meus amigos do "Crossbox Montelongo", em especial ao treinador Tiago Freitas, que mesmo nas alturas mais complicados ensinaram-me a nunca desistir.

À minha "família" de Braga que me acompanhou desde o início da universidade.

Um especial agradecimento ao CESIUM que me proporcionou momentos incríveis e inesquecíveis enquanto colaborador e mais tarde membro da direção. O CESIUM contribuiu para o meu crescimento como pessoa e tornou-me uma pessoa mais altruísta. Se hoje sou capaz de enfrentar novos desafios

em parte devo-o ao CESIUM. Muito obrigado a todos que comigo fizeram parte deste grande núcleo e boa sorte para as gerações vindouras.

Aos funcionários do Complexo Desportivo da Universidade do Minho, um obrigado pela paciência demostrada e por contribuírem para o meu crescimento ao longo destes anos.

Aos funcionários do Departamento Informática, em especial as senhoras da secretaria, um muito obrigado.

Aos meus colegas e amigos do "Hotella" obrigado por me ajudarem a crescer enquanto pessoa e por criarem um espaço onde foi possível cultivar o conhecimento e a aprendizagem.

Aos docentes da Universidade do Minho, em especial aos docentes do Departamento de Informática, que direta ou indiretamente contribuíram para a minha aprendizagem ao longo destes anos.

Ao Professor Doutor Paulo Carvalho pelo contributo de vital importância nas temáticas ligadas à área de rede envolvidas na minha dissertação.

Ao Professor Doutor Orlando Belo gostaria de lhe agradecer por tudo que fez e não me refiro apenas facto de me ter ajudado nesta dissertação, como também a tudo que fez pelos alunos onde se demonstrou sempre apto para ajudar os alunos nas mais várias formas possíveis. A sua basta experiência como orientador ajudou para que fosse possível cumprir os objetivos traçados para esta dissertação.

A todos aqueles que direta ou indiretamente me acompanharam neste caminho o meu profundo e sincero agradecimento.

Resumo

Process Mining na Análise de Tráfego de uma Rede de Comunicações

Nos primórdios a Internet era usada apenas por algumas pessoas. Nessa altura muitas das grandes empresas de tecnologia ainda não tinham aparecido. Porém tudo mudou quando a Internet entrou nos circuitos comerciais, provocando o aparecimento de estruturas para fazer a ligação das pessoas ao seu mundo. Desde aí que as grandes empresas têm adotado novas formas de encarar o mercado, aumentando gradualmente a sofisticação da forma de o fazerem. As empresas começaram a monitorizar a atividade dos seus clientes para que com isso melhorar a oferta dos seus bens e serviços ao público em geral. Todavia o conhecimento acerca daquilo que o cliente gosta não é suficiente para o atrair. Uma empresa também precisa que a informação apresentada ao cliente seja feita da maneira mais rápida possível. Por exemplo, se um cliente esperar mais do que "três" segundos para que o site seja carregado, o cliente irá abandoná-lo e, provavelmente, procurar um outro site de uma empresa concorrente. A Google avalia a rapidez dos sites e com isso dá-lhes uma pontuação. Os sites com piores pontuações são apresentados em últimos, o que tem, como sabemos, um grande impacto na escolha dos clientes. Mas não é só com os clientes que as empresas se tem de preocupar. Internamente os serviços dos funcionários de uma empresa podem ser afetados por uma Internet lenta, o que conduz a uma perda de performance e ao aumento da frustração do próprio funcionário no local de trabalho. Por estas razões é importante que as empresas estejam constantemente a monitorizar o tráfego passado pelos seus servidores, para serem capazes de verificar se os motivos da lentidão dos seus serviços

de rede são internos ou não. Neste trabalho de dissertação desenvolvemos um trabalho baseado em *process mining,* que através de uma ferramenta de monitorização de rede, *wireshark*, permite avaliar a qualidade de serviço da rede através da observação e análise das *logs* produzidas por alguns dos seus equipamentos, em particular, dos seus routers. Como são geradas várias *logs* para cada um tipo de router foi necessário fazer a sua conciliação, para que, a partir daí, se pudesse obter o percurso que os vários pacotes realizaram na sua movimentação pela rede. Desta forma, é possível criar um modelo matemático capaz de determinar um índice de bem-estar relativo à qualidade de serviço da rede de uma empresa. Basicamente, este índice permitirá avaliar o desempenho da rede e permitir aos seus gestores identificar, por exemplo, quais os pontos da rede que apresentam menor desempenho (ou estrangulamentos de serviço) e prevenir futuras quebras no serviço geral da rede em análise.

Palavras-chave: *Process Mining,* Ferramentas de análise de tráfego de rede, Índice de qualidade, *Smart Cities*, *data warehouse.*

Abstract

Process Mining for Traffic Analyses of a Communication Network

In the early days the Internet was only used by some people. By then many of the big technology companies had not yet appeared. But everything changed when the Internet entered the market, causing the appearance of structures to connect people to the world. Since then, large companies have adopted new ways of looking at the market. Companies have begun to monitor the activity of their customers so that they can improve the supply of their goods and services to the public. However, knowledge about what the client likes is not enough to attract him. A company also needs the information presented to the customer is made as quickly as possible. For example, if a client waits more than "three seconds" for the site to load, the client will abandon it and probably look for another site from a competing company. Google evaluates the speed of websites and gives them a score. The sites with the worst scores are presented in the last, which has, as we know, a great impact on the choice of customers. But it is not just with customers that companies must worry. Internally the services of a company's employees can be affected by a slow Internet, which leads to a loss of performance and to the frustration of the employee himself in the workplace. For these reasons it is important that companies are constantly monitoring traffic passed by their servers to be able to verify that the reasons for the slowness of their network services are internal or not. In this dissertation we developed a work based on process mining, which through a network monitoring tool, wireshark, allows to evaluate the quality of service of the network through the observation and analysis of logs produced by some of its equipment the routers. As

several logs are generated for each type of router, it was necessary to reconcile it, so we could obtain the route that the various packages made through the network. In this way, it is possible to create a mathematical model capable of determining an index of well-being related to the quality of service in a company's network. Basically, this index will allow assessing the performance of the network and allow its managers to identify, for example, which network points have the lowest performance (or service bottlenecks) and prevent future outages in the network.

Key-words: Process Mining, Network Monitoring and Analysis Tools, Index quality, Smart Cities, *data warehouse.*

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 Contextualização

Nos dias que correm a Internet desempenha um papel fundamental no nosso dia a dia. Com ela, é possível fazer tudo ou praticamente tudo sem sair do conforto da nossa casa. Isso revolucionou de forma muito acentuada a maneira como as empresas se posicionavam no mercado, bem como contribuiu para o surgimento de novas empresas tecnológicas, como são os casos do Facebook, da Amazon, ou da Google. Hoje, as empresas lutam entre si para aumentar o número de utilizadores que frequentam os seus sites. Para que isso aconteça elas sabem que é preciso melhorar o serviço e a oferta dos serviços e produtos que apresentam aos seus clientes.

A monitorização das atividades que clientes realizam nos sues Websites representa para as empresas uma grande fonte de informação. É através dessa monitorização que as empresas conseguem melhorar aquilo que oferecem aos seus clientes. As informações que vão sendo recolhidas incluem, entre outras coisas, a informação pessoal do próprio cliente, como a sua idade, a sua localização geográfica, o seu historio da internet, etc. É através desta informação e com a aplicação de algoritmos de *machine learning* que as empresas conseguem fazer sugestões de

serviços ou produtos de acordo com o perfil dos seus clientes e das suas necessidades. Toda esta informação tem que ser rapidamente apresentada aos vários utilizadores, para que estes não terminem as suas atividades no site e passem a visitar um outro qualquer, numa empresa da concorrência. Internamente também é necessário que as empresas apresentem uma boa qualidade de serviço no que diz respeito ao acesso e utilização da Internet. Por exemplo, um dos critérios que a Google usa para avaliar um site é a rapidez que este é carregado. Quanto mais lento for o carregamento de um site, menor cotação irá ter. A cotação de um site tem bastante impacto, visto que a Google apresenta os seus resultados de acordo com essa cotação. Isto significa que, quanto menor for essa cotação, pior será a localização do site em termos de qualidade de acesso. Como consequência, baseando-se nesta informação, um qualquer utilizador poderá não escolher esse site.

Uma fraca ligação à Internet poderá resultar numa perda de performance dos trabalhadores de uma empresa e mesmo até num aumento na sua frustração em termos de desempenho e qualidade de serviço. Um acesso lento à Internet poderá afetar uma empresa tanto internamente, no caso dos seus trabalhadores, como externamente, na aquisição de novos clientes ou na sua fidelização. Logo é necessário proceder à implementação de medidas que contrariem isso, como, por exemplo, fazer a monitorização dos seus servidores e identificar, o mais rapidamente possível, quais os pontos críticos do seu sistema de comunicações, da sua rede de computadores. Na prática, isto implica conhecer quais os locais onde os sistemas falham ou que existe a possibilidade de ocorrerem falhas num futuro próximo.

Para que seja possível detetar possíveis anomalias num sistema de rede de comunicações é necessário que o seu gestor de rede possuía um conhecimento alargado do sistema que está sob a sua alçada. Isto só é conseguido através de ferramentas de monitorização de rede, que permitem a recolha de informação pertinente sobre o tráfego da rede de comunicações, informando com detalhe e precisão aquilo que os seus diversos componentes estão ou não a fazer. Uma vez recolhida essa informação é necessário fazer a sua preparação e análise, que passa não raras vezes pela junção dos vários elementos de dados, usualmente armazenados em *logs* de serviço, e perceber a forma como as "coisas" acontecem no sistema de comunicações, incluindo conhecer o caminho percorrido por cada pacote no sistema de rede. Após esse trabalho estar realizado, através das técnicas de *process mining* é possível obter elementos bastante ilustrativos sobre o comportamento da rede e, inclusivamente, do caminho que cada pacote atravessa até chegar ao

seu destino. Depois, consoante os resultados desse processo de análise, poder-se-á desenvolver uma estrutura de dados adequada para análise-um *Data Warehouse* -, para acolher os diversos dados de monitorização e traçar um perfil temporal sobre os vários elementos constituintes da rede, mediante de um cálculo de um índice de bem-estar, de acordo com um conjunto de parâmetros previamente definidos que sustentem o cálculo desse índice com base nas características do sistema de rede e dos seus componentes. Tal índice irá permitirá identificar, por exemplo, quais os servidores que estão a causar atrasos ou quebras no sistema, o que contribuirá para definir estratégias adequadas para a sua rápida resolução, evitando assim índices de qualidade de serviço baixos que, como sabemos, podem trazer prejuízos para às próprias empresas.

1.2 Motivação

Uma rede de computadores empresarial deverá ser pensada, desenhada e implementada no sentido de dar uma resposta eficaz às necessidades do negócio (organização) que suporta, quer estas sejam de âmbito mais ligeiro, de acesso pouco frequente, quer estas advenham de grandes volumes de tráfego. Na realidade, as redes são, hoje em dia, a base de todo o trabalho desenvolvido dentro e fora de portas, tanto em termos de telecomunicações ou de ligação aos servidores, como em termos do funcionamento dos sistemas internos ou de outras necessidades diárias que as empresas possam ter.

Atualmente, as infraestruturas de uma rede são de alta tecnologia, envolvendo processos de comunicação de alta velocidade para que possam responder, de forma eficaz, sempre e quando seja necessário às diversas solicitações que os processos de comunicação lhes requerem. A verdade é que, cada vez mais, as empresas precisam de ter um sistema de comunicações fiável e robusto para garantirem que a sua regular operação e a sua produtividade estejam ao mais alto nível. Assim, é importante assegurar a capacidade de um sistema de comunicações para executar a sua função de forma contínua (sem interrupção), por um período de tempo significativo e com uma boa qualidade de serviço. O importante é, pois, assegurar que todo o equipamento relacionado com o sistema de comunicações, por exemplo os servidores ou os routers, estejam permanentemente disponíveis, independentemente do dia, da hora, do local ou de outros fatores

que possam influenciar a sua disponibilidade em determinado momento. Para que tal aconteça é necessário que as empresas estejam munidas de ferramentas que permitem identificar os problemas do seu sistema de comunicações, de forma rápida e eficaz, bem como saber identificar as causas dos problemas que eventualmente ocorram.

1.3 Objetivos

No âmbito deste trabalho de dissertação estabeleceu-se um objetivo bastante claro: identificar quais os pontos de estrangulamento de uma rede de comunicações. Para concretizar tal objetivo, desenvolvemos um processo com vista à verificação do desempenho da rede, através da definição e manutenção de um índice de desempenho, para cada ponto da referida rede. O processo referido foi idealizado de forma a poder correlacionar os aspetos mais relevantes que condicionam o desempenho do sistema de rede, através da análise do funcionamento de cada um dos seus pontos. Na prática, isto significa que, se fez a análise do volume de dados que circula em cada um desses pontos, dos pacotes perdidos nesses pontos, da velocidade de comunicação, do tempo que demora até estabelecer a ligação e do tempo que demora a processar os pacotes recebidos. Dessa forma foi possível monitorizar a rede, ponto a ponto, o que permitiu obter uma informação sobre o funcionamento da rede bastante mais detalhada e rigorosa. Com isso conseguimos avaliar o desempenho do sistema de comunicação e determinar a prior potenciais pontos de estrangulamento do sistema, o que nos permite desenvolver atempadamente medidas para evitar possíveis quebras de serviço. Para sustentar todo este processo de recolha de dados e de monitorização de funcionamento do sistema de comunicações concebemos e implementámos também um sistema de *dashboarding*, iterativo, que permita verificar, a cada momento, o estado da rede (Figura 1).

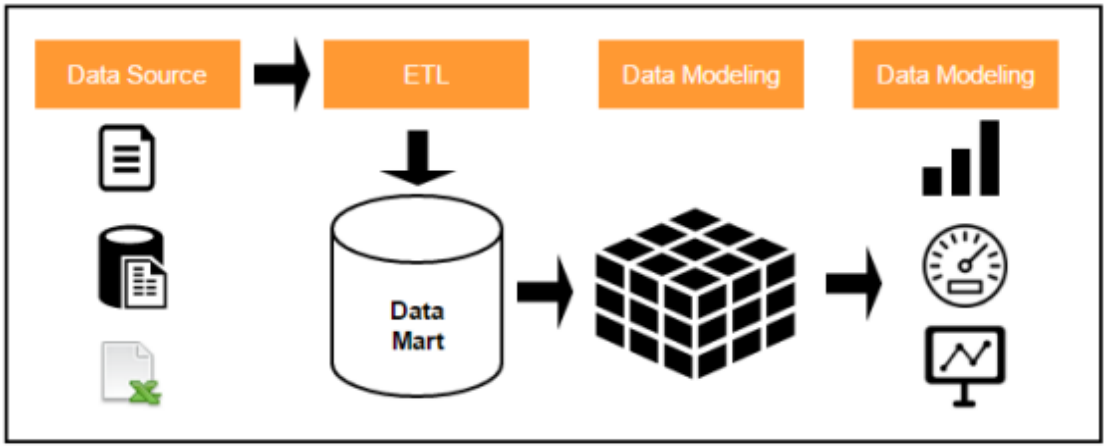


Figura 1-Esquema ilustrativo da sequência dos processos que foram implementados.

De uma forma sucinta, podemos dizer que o trabalho realizado no âmbito desta dissertação, contribuiu para:

- Desenvolver um método que nos permitisse a cada momento conhecer o nível de desempenho (a qualidade de serviço) de uma rede de comunicações por computador.
- Desenvolver um conjunto de mecanismos que facilitasse a interação entre o utilizador e o sistema.
- Desenvolver uma estrutura responsável pelo tratamento e armazenamento dos dados recolhidos de uma rede de comunicações.
- Desenvolver uma estrutura de suporte que permitisse ao utilizador retirar conclusões sobre o real estado de uma rede de comunicações.

1.4 Organização do Documento

Para além do presente capítulo, esta dissertação está organizada da seguinte maneira:

- Capítulo 2 - Um Caso de Estudo - neste capítulo apresentamos a rede de comunicação que idealizámos, as ferramentas usadas na monitorização de rede, as ferramentas usadas nesta dissertação para a monitorização de rede e os ficheiros *log* resultados dessa monitorização.

- Capítulo 3 - O Estudo da Rede - neste capítulo abordamos o que é o *process mining*, para que é utilizado, as suas vantagens e as ferramentas usadas. Abordamos também a forma como tratamos os dados recolhidos com a monitorização da rede de comunicação, a fim de ser possível aplicar algoritmos de *process mining*. Por fim mostramos uma rede gerada pelo algoritmo de *process mining*.
- Capítulo 4 - Os Dados do Tráfego Gerado - neste capítulo começamos por analisar os dados utilizados pelo algoritmo de *process mining*. Esta análise consiste em perceber a natureza dos dados, se são numéricos, alfabéticos ou datas, perceber também quais os atributos que serão guardados e quias os que serão descartados a fim de serem guardados num sistema que irá servir de suporte à decisão. Posto isto, os dados serão carregados para uma estrutura ETL, para proceder ao respetivo tratamento. Posto este tratamento, os dados serão carregados para um *data mart* idealizado. Terminámos com a apresentação do modelo da base de dados multidimensional que será implementada no sentido de agilizar todo o processo de análise das métricas e indicadores recolhidos.
- Capítulo 5 - Um Índice de Monitorização - neste capítulo apresentamos a expressão usado no cálculo do índice de bem-estar, bem como uma explicação sobre os atributos usados no cálculo do índice e terminámos com uma explicação das infraestruturas usadas- ETL, *data warehouse* e base de dados multidimensional.
- Capítulo 6 - Análise do Sistema de Monitorização - neste capítulo começamos por introduzir uma breve explicação sobre o que é um *dashboard*, para que é usado, assim como os vários tipos de *dashboard*. No fim apresentamos o *dashboard* final onde são visíveis estes indicadores.
- Capítulo 7 - Conclusões e Trabalho Futuro - neste capítulo apresentamos algumas conclusões sobre este trabalho de dissertação, analisando aspetos positivos e negativos que o tenham envolvido. Concluímos o capítulo com uma breve análise relativamente ao futuro do método desenvolvido e da sua aplicação em ambientes reais.

Capítulo 2

Um Caso de Estudo

2.1 Estabelecimento da entidade do projeto

A universidade *Universitas*[1] é uma universidade muito prestigiada, tanto a nível nacional como internacional, que oferece aos seus alunos um basto leque de cursos. Devido ao seus alunos e professores se terem destacado em várias áreas do conhecimento, tendo tido reconhecimento tanto a nível nacional como internacional, fez com que houvesse um aumento das vagas disponíveis para admissão à universidade, o que contribuiu para a emergência de alguns novos problemas para universidade. Um desses problemas foi o decréscimo da qualidade de serviço do seu acesso à Internet, que, basicamente, ficaram mais lentos. Consequentemente, os alunos começaram a queixar-se que a Internet estava sempre ir abaixo e que havia pontos no campus onde simplesmente não havia acesso à Internet. Como tal, a universidade decidiu melhorar os serviços de acesso à Internet aos alunos. Todavia, nesta altura, a universidade não sabe o que deve ser, de facto, melhorado e onde deve ser melhorado.

[1] A universidade Universitas não existe e foi criada exclusivamente como caso de estudo para esta dissertação.

2.2 A rede de comunicação idealizada

Nesta altura é importante revelar o nosso objeto de estudo, a rede de comunicações por computador que vamos utilizar para fazer o estudo e caracterização de comportamentos anómalos ou indesejáveis. O sistema de rede da universidade que idealizámos pode ser observado na Figura 2. Os vários equipamentos -computadores, routers, etc.- que nela figuram representam as várias escolas, complexos pedagógicos, bibliotecas e outros edifícios que incorporam a estrutura da universidade. Podemos imaginar que, dentro desses edifícios, os serviços do sistema de rede em questão são utilizados por alunos, docentes, não-docentes, investigadores, entre outros, através das suas plataformas computacionais, tentar aceder à Internet.

A rede idealizada contém vários equipamentos, todos com várias ligações, de saída ou de entrada, para que não seja percetível o caminho que os vários pacotes seguem, assim como, não seja possível perceber a priori quais os pontos problemáticos, isto é, pontos que não estejam a funcionar como deveriam. É importante que a rede se assemelhe o mais possível à realidade, por isso é preciso idealizar uma rede de modo a permitir que vários pacotes circulem nos vários pontos intermédios antes de chegar ao seu destino final. Ao longo da rede as ligações apresentaram a mesma velocidade de transporte, 512 kilobit por segundo (*kbps*), que mediante dos resultados apresentados poderá sofrer alterações em alguns pontos da rede idealizada.

Se analisarmos a configuração apresentada podemos ver, por exemplo que, o router *n19* (na figura, o elemento da rede posicionado mais à direita) é aquele que garante o acesso da universidade ao exterior. Como tal, todos os outros routers da rede que queiram comunicar com o exterior têm de encaminhar o seu tráfego para este router. A rede interna idealizada foi criada usando a ferramenta *CORE network emulator* [20].

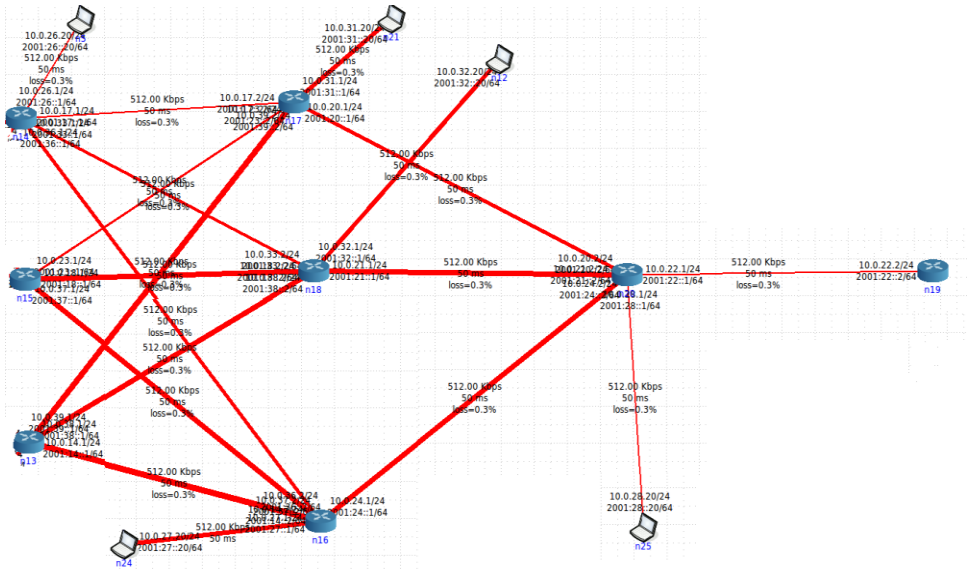


Figura 2-Representação parcial da rede interna da *Universitas.*

2.3 Monitorização da rede

Para fazermos a monitorização da rede usámos uma ferramenta específica para o efeito: *wireshark* [1]. A partir dela conseguimos recolher informação acerca de cada ligação associada a um router[2], bem como o tráfego proveniente dos diversos "computadores" clientes da rede. Para simular a comunicação entre os vários utilizadores da rede, registámos o comportamento da rede através da execução sucessiva de comandos: *ping -i tempoEntreAsTramas*[3] *destino*[4]. Este comando só foi utilizado para o caso particular dos computadores, uma vez que serão eles os responsáveis pela geração do tráfego da rede (Figura 3).

[2] Quando um router está ligado a um outro router a ferramenta cria uma nova ligação. Por isso se um router comunicar com três outros routers a ferramenta permite fazer a recolha de dados dessas três ligações.
[3] Intervalo de tempo entre cada trama enviada. Exemplo ping -i 0.001 10.0.22.2. Neste caso o intervalo de tempo em cada trama enviada é de 1 ms.
[4] Endereço de IP onde será enviada a trama, no nosso caso é o endereço 10.0.22.2.

Para definirmos uma base de trabalho coerente e representativa do comportamento da rede que idealizámos, fizemos a geração de tráfego para cada um dos dias da semana, de segunda a domingo, num intervalo de tempo de seis horas.

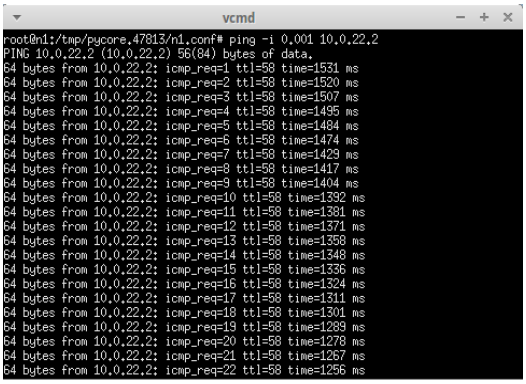


Figura 3-Ilustração do resultado da aplicação do comando usado para a geração do tráfego da rede.

Como o *wireshark* capa muitas tramas que não pretendemos analisar, tivemos que filtrar alguma dessa informação. Para isso usámos o seguinte comando *ip.dst==10.0.22.2.* que nos permitiu filtrar apenas as tramas que tiveram como destino o endereço de IP 10.0.22.2, descartando-se assim todas as tramas que não tenham como destino o IP 10.0.22.2 (Figura 4).

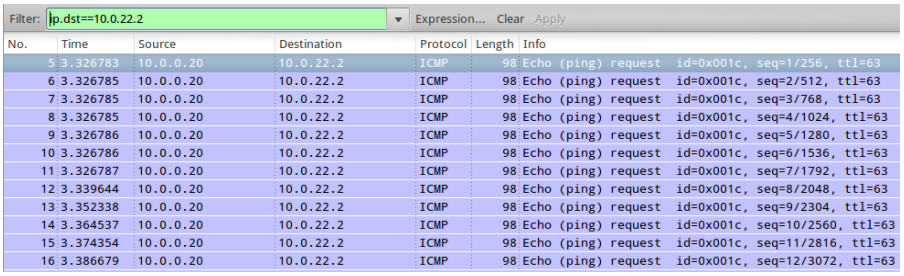


Figura 4-Exemplo de uma captura de dados após a aplicação do filtro *ip.dst==10.0.22.2.*

Na figura 5 está apresentado um caso no qual nenhuma trama que circula naquela ligação tem como destino o endereço de IP 10.0.22.2.

Figura 5-Exemplo de uma captura na qual o filtro não é satisfeito.

2.4 O Registo do tráfego da rede

Após a execução do processo de geração do tráfego da rede idealizada, verificámos os registos efetuados pela ferramenta de monitorização relativos a cada ligação do router e dos "computadores" envolvidos nos processos de comunicação registados. De seguida, procedemos à sua combinação, uma vez que é necessário reuni-los num único ficheiro e, depois, criar uma rede na qual seja possível observar o caminho percorrido pelos vários pacotes ao longo dos seus diversos nodos. Para isso, desenvolvemos um programa em *JAVA* especialmente orientado para a junção dos vários ficheiros de registo de tráfego e criação de um único ficheiro de *log* contendo toda a informação recolhida acerca dos processos de comunicação na rede. Na Figura 6 podemos ver um fragmento da *log* final após a aplicação do programa referido.

casos	source	idTrama	seq	ipOrigem	tamanho	tempo
1	10.0.34.20	27	20417/49487	10.0.34.20	784	2018-05-26 12:40:37
1	10.0.14.1	27	20417/49487	10.0.34.20	784	2018-05-26 13:07:14
1	10.0.24.1	27	20417/49487	10.0.34.20	784	2018-05-26 13:33:51
1	10.0.22.1	27	20417/49487	10.0.34.20	784	2018-05-26 14:00:28
1	10.0.22.2	27	20417/49487	10.0.34.20	784	2018-05-26 14:27:05
2	10.0.28.20	35	32924/40064	10.0.28.20	784	2018-05-26 12:51:55
2	10.0.22.1	35	32924/40064	10.0.28.20	784	2018-05-26 13:29:50
2	10.0.22.2	35	32924/40064	10.0.28.20	784	2018-05-26 14:07:45
3	10.0.40.20	27	2565/1290	10.0.40.20	784	2018-05-26 12:26:16
3	10.0.37.1	27	2565/1290	10.0.40.20	784	2018-05-26 12:38:32
3	10.0.24.1	27	2565/1290	10.0.40.20	784	2018-05-26 12:50:48
3	10.0.22.1	27	2565/1290	10.0.40.20	784	2018-05-26 13:03:04
3	10.0.22.2	27	2565/1290	10.0.40.20	784	2018-05-26 13:15:20
4	10.0.40.20	27	2565/1290	10.0.40.20	784	2018-05-26 13:22:25
4	10.0.37.1	27	2565/1290	10.0.40.20	784	2018-05-26 14:30:50
4	10.0.24.1	27	2565/1290	10.0.40.20	784	2018-05-26 15:39:15
4	10.0.22.1	27	2565/1290	10.0.40.20	784	2018-05-26 16:47:40
4	10.0.22.2	27	2565/1290	10.0.40.20	784	2018-05-26 17:56:05
5	10.0.28.20	35	3266/49676	10.0.28.20	784	2018-05-26 12:25:54
5	10.0.22.1	35	3266/49676	10.0.28.20	784	2018-05-26 12:37:48
5	10.0.22.2	35	3266/49676	10.0.28.20	784	2018-05-26 12:49:42
6	10.0.31.20	27	12015/61230	10.0.31.20	784	2018-05-26 12:34:30
6	10.0.20.1	27	12015/61230	10.0.31.20	784	2018-05-26 12:55:00
6	10.0.22.1	27	12015/61230	10.0.31.20	784	2018-05-26 13:15:30
6	10.0.22.2	27	12015/61230	10.0.31.20	784	2018-05-26 13:36:00
7	10.0.28.20	35	50035/29635	10.0.28.20	784	2018-05-26 13:06:56
7	10.0.22.1	35	50035/29635	10.0.28.20	784	2018-05-26 13:59:52
7	10.0.22.2	35	50035/29635	10.0.28.20	784	2018-05-26 14:52:48

Figura 6-Fragmento do ficheiro *Log* final com as comunicações estabelecidas.

2.5 Monitorização do tráfego da rede

O processo de monitorização rede não serve apenas para detetar se a rede está lenta ou não, mas também irá ser útil, para verificar se a rede utilizada é segura ou não, ou se está sendo vítima de algum ataque de terceiros. Independentemente do tipo de monitorização usado é preciso ter em conta alguns outros aspetos, como o tipo de equipamento, por exemplo, usado também terá influência na velocidade da rede. Existem variados fatores que podem influenciar o desempenho de uma rede. Por isso é que, sempre que uma empresa opte pela sua monitorização tenha em conta os diferentes fatores e não apenas a velocidade da rede.

Existem dois tipos de monitorização: a monitorização interna, onde é a própria entidade a fazer essa monitorização, através das várias ferramentas disponíveis [21], ou então a monitorização externa, onde contratam uma empresa [22] para fazer esse tipo de monitorização. A informação que se pretende retirar da rede também irá influenciar o tipo de ferramenta que a utilizar. Se se pretender utilizar uma ferramenta que analise não só o tráfego da rede como também a própria segurança do sistema de comunicações, então ter-se-á de optar, por exemplo, por uma ferramenta como o *GFI LanGuard* [23] (Figura 7). Caso, se se pretender, apenas, perceber a quantidade de informação que circula num determinado equipamento, então uma ferramenta como o *Wireshark* (Figura 8) serve perfeitamente. É importante realçar que, antes de qualquer monitorização a realizar se saiba que tipo de análises se pretende realizar sobre a rede e que informação precisamos para suportar adequadamente tais processos.

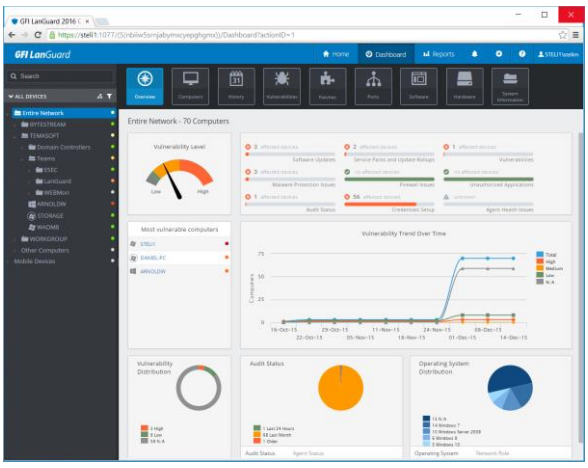


Figura 7-Interface da ferramenta *GFI LanGuard.*

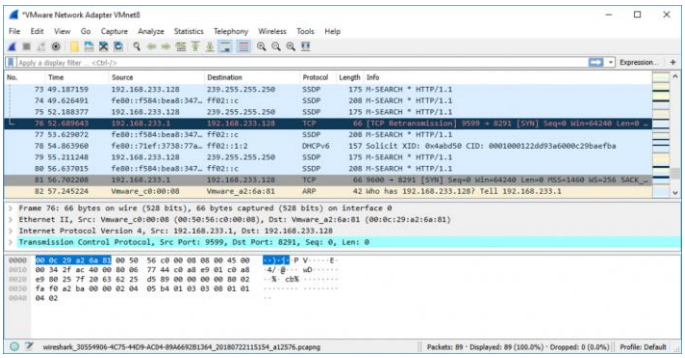


Figura 8- Interface da ferramenta *Wireshark.*

Embora estas ferramentas sejam bastantes úteis e poderosas, todavia não permitem ver a maneira como os pacotes são percorridos na sua rede. Para isso temos que recorrer a outro tipo de ferramenta mias especializada e orientada particularmente para a análise de informação contida em ficheiros *log* de eventos. A utilização de técnicas de *process mining*, permite-nos alcançar esse patamar de análise, uma vez não só permitem detetar potenciais situações anómalas na rede, através da análise dos registos de comunicação contidos nos ficheiros *log*, como também permitem ao utilizador perceber como viajam os pacotes na própria rede. Este processo de análise será apresentado e discutido no capítulo seguinte.

Capítulo 3

O Estudo da Rede

3.1 Mineração de Processos

Hoje em dia as empresas recolhem muita informação dos seus clientes. Porém, mais importante do que recolher simplesmente a informação é extrair conhecimento útil dessa mesma informação. É esse o objetivo de técnicas de extração de conhecimento como a mineração de possessos (*process mining*) [2]. Em particular, a mineração de processos atua sobre a informação recolhida acerca da realização de processos, analisando os diversos eventos ocorridos, e a partir daí apresentar modelos de análise de fácil interpretação sobre o comportamento registado. A mineração de processos é uma técnica relativamente recente que assenta entre modelos e técnicas de *machine learning* e *data mining,* por um lado, e por outro entre técnicas e modelos de *process modeling* e *analysis*. Basicamente, quando se utiliza uma técnica de mineração de processos tem-se como objetivo descobrir, monitorizar e melhorar algum tipo de processos, através da análise do seu comportamento (da sua execução) ao longo de um ou mais períodos de execução [2].

Usualmente, os modelos criados através de *process mining* têm um grande impacto nas organizações, seja para formular ou discutir ideias (*informal models*) como também para retirar conhecimento de uma determinada da rede (*formal models*). Antes da criação de um ou vários modelos é necessário analisar os ficheiros *logs* com os eventos registados. Sem uma análise

cuidada desses ficheiros o modelo gerado poderá induzir em erro quem o está a analisar, e como tal podendo ter um impacto muito negativo no futuro. Uma das grandes dificuldades do *process mining* é a obtenção de dos ficheiros de *log*, da informação de trabalho. Apesar das empresas registarem muita informação sobre os seus diversos trabalhos, nem sempre é fácil aplicar os algoritmos de *process mining* sobre essa informação.

As técnicas de *process mining* podem ser aplicadas em várias áreas de aplicação. Exemplos destas áreas onde podem ser aplicados estas técnicas de *process mining* são: os transportes, a saúde, a banca, as telecomunicações, entre outras [24]. Em seguida apresentamos alguns exemplos reais nos quais foi utilizado *process mining* para melhorar procedimentos já existentes, nomeadamente:

- Uma companhia holandesa, *Nederlandse Spoorwegen* [25], fez um estudo sobre o aluguer das suas bicicletas (*OV-Bike*), e descobriu que as bicicletas que foram dadas como roubadas, afinal tinham sido entregues. Este erro fez com que a empresa gastasse mais dinheiro na compra de novas bicicletas [26].
- Um hospital universitário holandês, *Academic Medical Center* [27], fez um estudo sobre as várias etapas percorridas pelos seus doentes [28]. Este estudo envolveu a realização de três processos de análise envolvendo, os trajetos percorridos pelos vários doentes enquanto se encontram no hospital, a comunicação existente entre os vários departamentos e a perspetiva de melhoramento. Com este estudo foi possível concluir que o *process mining* pode ser uma ajuda preciosa para ajudar os hospitais a melhorar o tratamento dos seus pacientes.
- A Vodafone [29] em parceria com a *Celonis* [30], analisou os dados presentes na *SAP HANA* [31] e com isso conseguiu melhorar o desempenho dos seus colaboradores, resultando num impacto positivo para a empresa [32].

3.2 Ferramentas de *Process Mining*

O process mining é uma área que nos últimos anos teve grande expansão, dadas as suas potencialidades para a descoberta de ineficiências em processos de natureza diversa. Hoje em dia no mercado existem algumas ferramentas que ajudam bastante na criação de redes de execução de processos, recorrendo a algoritmos bem conhecidos. Dentro do domínio do *process mining* as duas ferramentas mais conhecidas são o *ProM* [3] e o *Disco* [4]. A primeira é uma ferramenta

open source, bastante poderosa, que disponibiliza vários algoritmos de *process mining*. É bastante útil para traçar redes de execução de processos. Apesar de o *ProM* ser bastante utilizado exige ao seu utilizador um conhecimento profundo sobre própria ferramenta, o que torna complicado a sua exploração e consequente análise dos resultados que disponibiliza. A ferramenta *Disco*, por seu lado, é mais fácil de utilizar. Apesar de não ser tão sofisticada como o *ProM*, esta ferramenta consegue gerar uma rede de processos de forma mais simples e de mais fácil interpretação. As duas ferramentas dão para ser utilizadas em conjunto, uma vez que o *ProM* só lê um tipo de formato de ficheiros de dados (*XES*[5]). Por isso recorremos à ferramenta *Disco* para converter um ficheiro *CSV* nesse formato e só depois é que processámos os nossos registos de eventos em *ProM.* Na Figura 9 é possível observar a realização do primeiro passo na aplicação de qualquer algoritmo de *process mining*, a escolha dos atributos a usar. Na Figura 10 é possível observar, depois de escolhido os atributos, os vários tipos de algoritmos que podemos usar sobre os dados.

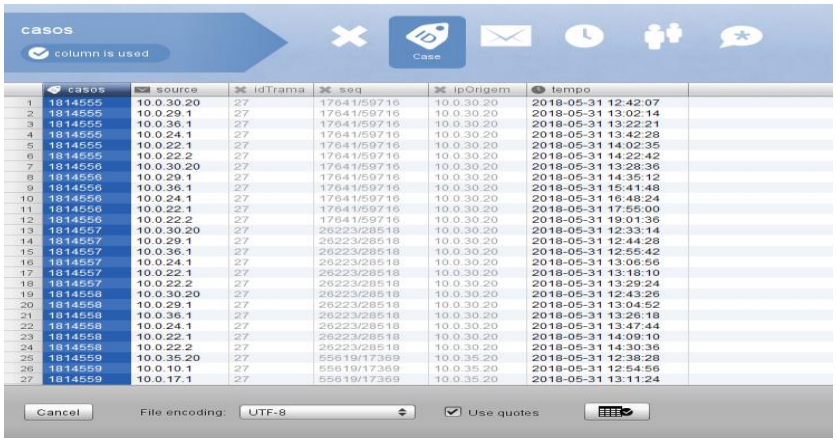


Figura 9-Uma vista da interface da ferramenta *Disco*.

[5] *eXtensible Event Stream (XES)* é o formato universal em *process mining*. O seu antecessor era o formato MXML (*Mining eXtensible Markup Language).*

Figura 10-Uma vista da interface da ferramenta *ProM*.

3.3 Os ficheiros *Log* do tráfego gerado

Para serem usados em algoritmos de *process mining*, os ficheiros de *logs* têm de estar ordenados por casos. Qualquer evento que possa ser relacionado com um caso é uma atividade. Posto isto, for necessário preparar os nossos ficheiros de *log* de acordo com os requisitos de processamento da ferramenta. Uma vez preparados os ficheiros, apenas temos que escolher qual a ferramentas a utilizar e proceder à criação da rede de eventos. Na Figura 11 podemos ver um exemplo de um dos ficheiros de *log* utilizado para criar uma rede capaz de representar os caminhos que seguiram os dos vários pacotes envolvidos nos processos de comunicação.

Na Tabela 1 estão apresentados os metadados relativos aos ficheiros de log que foram utilizados nos nossos processos de geração de rede de eventos e de análise. Porém, os campos "*idTrama", "seq", "ipOrigem" e "tamanho"* não serão utilizadas pelos algoritmos de *process mining* na criação da rede. As colunas "*idTrama", "seq"* e *"ipOrigem"* foram utilizadas apenas no processo de identificação de cada pacote que foi gerado, de forma a que fosse possível traçar os diversos caminhos que os pacotes percorreram.

casos	source	idTrama	seq	ipOrigem	tamanho	tempo
1	10.0.34.20	27	20417/49487	10.0.34.20	784	2018-05-26 12:40:37
1	10.0.14.1	27	20417/49487	10.0.34.20	784	2018-05-26 13:07:14
1	10.0.24.1	27	20417/49487	10.0.34.20	784	2018-05-26 13:33:51
1	10.0.22.1	27	20417/49487	10.0.34.20	784	2018-05-26 14:00:28
1	10.0.22.2	27	20417/49487	10.0.34.20	784	2018-05-26 14:27:05
2	10.0.28.20	35	32924/40064	10.0.28.20	784	2018-05-26 12:51:55
2	10.0.22.1	35	32924/40064	10.0.28.20	784	2018-05-26 13:29:50
2	10.0.22.2	35	32924/40064	10.0.28.20	784	2018-05-26 14:07:45
3	10.0.40.20	27	2565/1290	10.0.40.20	784	2018-05-26 12:26:16
3	10.0.37.1	27	2565/1290	10.0.40.20	784	2018-05-26 12:38:32
3	10.0.24.1	27	2565/1290	10.0.40.20	784	2018-05-26 12:50:48
3	10.0.22.1	27	2565/1290	10.0.40.20	784	2018-05-26 13:03:04
3	10.0.22.2	27	2565/1290	10.0.40.20	784	2018-05-26 13:15:20
4	10.0.40.20	27	2565/1290	10.0.40.20	784	2018-05-26 13:22:25
4	10.0.37.1	27	2565/1290	10.0.40.20	784	2018-05-26 14:30:50
4	10.0.24.1	27	2565/1290	10.0.40.20	784	2018-05-26 15:39:15
4	10.0.22.1	27	2565/1290	10.0.40.20	784	2018-05-26 16:47:40
4	10.0.22.2	27	2565/1290	10.0.40.20	784	2018-05-26 17:56:05
5	10.0.28.20	35	3266/49676	10.0.28.20	784	2018-05-26 12:25:54
5	10.0.22.1	35	3266/49676	10.0.28.20	784	2018-05-26 12:37:48
5	10.0.22.2	35	3266/49676	10.0.28.20	784	2018-05-26 12:49:42
6	10.0.31.20	27	12015/61230	10.0.31.20	784	2018-05-26 12:34:30
6	10.0.20.1	27	12015/61230	10.0.31.20	784	2018-05-26 12:55:00
6	10.0.22.1	27	12015/61230	10.0.31.20	784	2018-05-26 13:15:30
6	10.0.22.2	27	12015/61230	10.0.31.20	784	2018-05-26 13:36:00
7	10.0.28.20	35	50035/29635	10.0.28.20	784	2018-05-26 13:06:56
7	10.0.22.1	35	50035/29635	10.0.28.20	784	2018-05-26 13:59:52
7	10.0.22.2	35	50035/29635	10.0.28.20	784	2018-05-26 14:52:48

Figura 11- Pequeno fragmento de um ficheiro *log*

Tabela 1-Metadados dos ficheiros de *log* utilizados

Colunas	Explicação
casos	A coluna "casos" representa o número diferente de pacotes que circulam na rede.
source	A coluna "source" representa o IP das várias ligações por onde os vários pacotes circularam.
idTrama	A coluna "idTrama" é um valor que identifica um pacote.
seq	A coluna "seq" é um valor que identifica um pacote.
ipOrigem	A coluna "ipOrigem" representa o local onde cada pacote se originou.
tamanho	A coluna "tamanho" é o tamanho em *bits* de cada pacote que passa naquela ligação.
tempo	A coluna "tempo" representa o instante em que o pacote passou por aquela ligação.

3.4 Análise da Rede Gerada

Depois do processo de obtenção dos dados ter sido realizado, procedemos a uma análise cuidada aos dados angariados. Neste processo de análise tivemos em conta o domínio dos dados, isto é, se os dados seriam numéricos, alfanumérico ou datas, entre outros. Um outro aspeto que se teve em conta neste processo de análise foi verificar se os eventos estavam ordenados de forma crescente

por data e por hora. Verificámos também se os vários eventos, presentes nos dados, estavam iniciados pelo IP que estava presente no campo "*ipOrigem*".

De seguida, fizemos a escolha da ferramenta de *process mining*, o Disco, de acordo com o exposto anteriormente, para fazermos a geração da rede de eventos baseada nos dados angariados sobre o funcionamento da rede. Na Figura 12 podemos ver rede[6] que foi gerada para um determinado dia específico, tendo em conta o número de pacotes que circularam entre cada ligação.

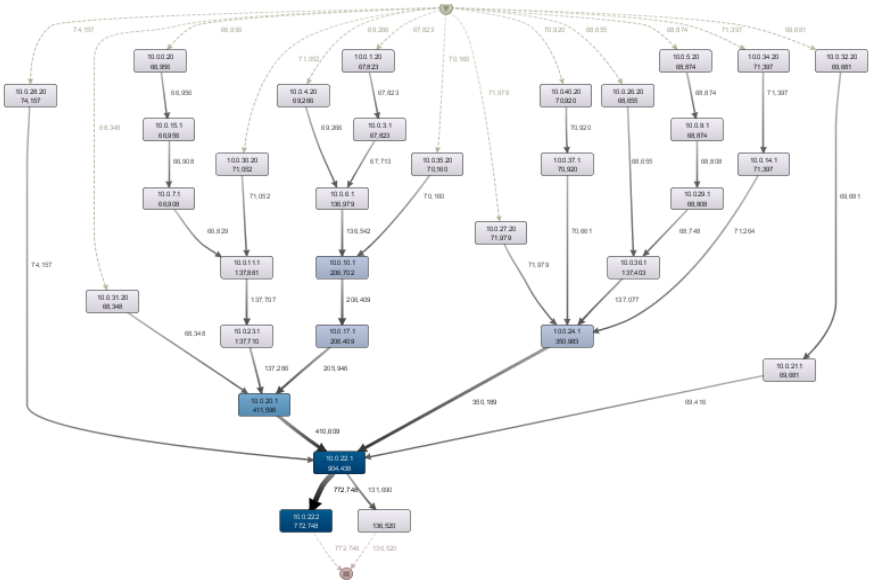


Figura 12-Exemplo de uma das redes gerada pela ferramenta de análise.

A rede gerada permite-nos ter uma ideia bastante concreta sobre o percurso efetuado pelos diversos pacotes no dia referido.

A Figura 13 apresenta um exemplo de um caminho gerado pela ferramenta de análise. Neste exemplo é possível observar os diferentes esquemas de cores, tendo em conta o número de pacotes que circulam nos pontos representados. À medida que o número de pacotes vai aumentando, a tonalidade da cor azul aumenta também. Neste exemplo é possível observar que os pacotes que tiveram origem no endereço de IP 10.0.27.20, foram depois encaminhados para o

endereço de IP 10.0.24.1. Neste encaminhamento não se verificou nenhum pacote perdido. No IP destino os pacotes foram depois encaminhados para o endereço de IP 10.0.22.1 ou perdidos na rede. No exemplo em questão rapidamente verificamos que 794 pacotes foram perdidos e 350189 foram encaminhados com sucesso. Posteriormente, os pacotes foram encaminhados para o endereço de IP 10.0.22.2, o que significa que tudo correu como esperado, ou ficaram perdidos na rede.

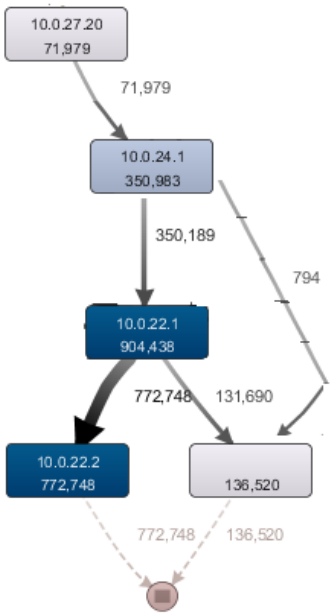


Figura 13- Exemplo de um caminho gerado pela ferramenta de análise em termos de frequência.

Através da ferramenta *Disco* foi também é possível identificar *outliers*, ou seja, registos que não seguiram o caminho que supostamente deveriam ter seguido. Porém, a observação da rede não é suficiente para retirar conclusões, porque a rede não tem em conta o funcionamento da rede nos restantes dias da semana. Como tal, esse processo de observação, isoladamente, pode contribuir para obtermos algumas conclusões precipitadas. A observação da rede é apenas um mero instrumento que ajuda na análise do problema. Antes de se retirar qualquer conclusão é preciso ter em conta os vários dias, bem como outros fatores que não estão presentes na rede. Esses

[6] O *Disco* permite ao utilizador encurtar os caminhos encontrados pelos algoritmos de *process mining*, sendo que a imagem

fatores, incluem o tipo de routers utilizados, por vezes os equipamentos falham e ao serem substituídos por um outro completamente diferente, podem causar uma discrepância na análise da rede, problemas relacionados com o distribuidor da internet, falhas de energia, entre outros fatores.

A ferramenta *Disco* também permite fazer uma análise em termos de performance da rede. Nesta vertente de análise podemos escolher o tipo de métricas que queremos usar, e.g. média, mediana, tempo total, tempo mínimo ou tempo máximo. No exemplo apresentado na Figura 14 podemos analisar a rede tendo em conta o valor médio que demora a enviar um dado pacote.

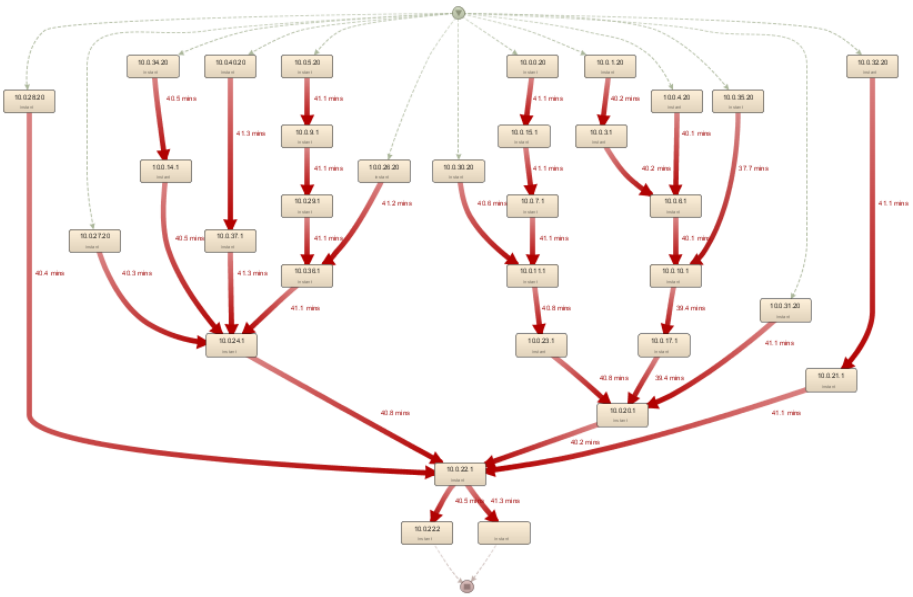


Figura 14-Exemplo de uma rede gerada tendo em conta performance de cada ligação.

A Figura 15 apresenta um exemplo de um caminho tendo em conta a performance do sistema, com origem no endereço de IP 10.0.27.20, onde é possível verificar o tempo médio que cada pacote demora a transitar de um sítio para o outro. Neste exemplo os tempos médios dos pacotes perdidos não interessam para análise da rede e serão descartados posteriormente.

apresentada é apenas uma amostra da rede original, não apresentando todos os caminhos presentes nos dados.

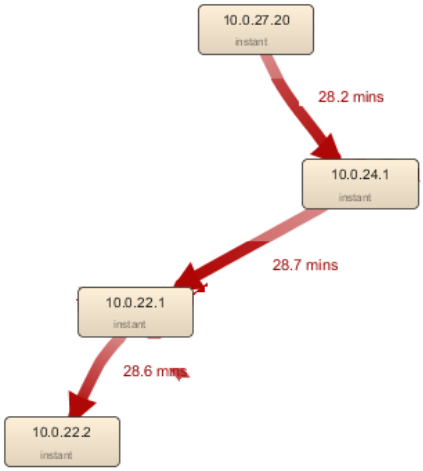


Figura 15- Exemplo de um caminho gerado pela ferramenta de análise em termos de performance.

Capítulo 4

Os Dados do Tráfego Gerado

4.1 O Sistema de Dados

Ao longo de qualquer processo de modelação é necessário conceber e desenvolver as estruturas e as ações necessárias que o modelo deverá refletir. Isso implica que se tenha um bom conhecimento acerca do problema em causa. Uma das partes mais relevantes do processo de modelação está relacionada com a criação da representação do sistema de dados que o modelo incorpora, o que implica uma compreensão da estrutura, do conteúdo, das relações e derivações dos dados. Como tal, é necessário verificar se os dados que temos à nossa disposição estão num estado utilizável ou se as suas falhas podem ser geridas.

O *Data profiling* utiliza recursos de consulta para explorar o conteúdo real e as relações que de alguma forma estão estabelecidas, explicita ou implicitamente, no sistema de dados, o que ajuda imenso na compreensão do sistema de dados que temos em mãos. Um processo de *Data profiling* pode ser tão simples como escrever algumas instruções em SQL ou tão sofisticado como envolver a utilização de uma ferramenta especializadas, como é o caso do *Data Cleaner* [35].

No caso particular deste trabalho dissertação, para garantirmos que os futuros processos de análise, tanto do volume de tráfego da rede como da sua qualidade de serviço, pudessem ser

concretizados de forma efetiva, desenvolvemos um processo cuidadoso para tratar do armazenamento dos dados recolhidos sobre o tráfego da rede, com o objetivo de preparar um sistema de dados adequado ao suporte dos processos de análise requeridos. Analisando a fonte de dados que temos à nossa disposição, facilmente verificámos que em certos pacotes não chegam ao destino, ou seja, esses pacotes são perdidos na rede. Isso transparece de imediato quando verificamos que o ponto de destino está, simplesmente, vazio. De qualquer forma, dado que os dados foram sintetizados em laboratório, problemas como esse são fáceis de ultrapassar. Assim, tendo em consideração situações como aquela que acabámos de descrever, procedemos à elaboração do mapeamento lógico dos dados, tendo em conta a fonte de dados referida e a estrutura de dados multidimensional, que fomos desenvolvendo em paralelo com o processo de análise da fonte de dados. Essa estrutura está incluída e suportada pela matriz de decisão apresentada na Tabela 2.

Tabela 2- Mapeamento lógico dos dados fonte-destino.

Target				source			Transfomation
database	table	table type	column	database	column	datatype	
dw_Trafego	DimCasos	dimension	casos	dia*.csv	casos	INT	Mudar o valor dos casos repetidos referentes ao vários dias
dw_Trafego	DimCasos	dimension	idCasos				Surrogate Key
dw_Trafego	DimLigacao	dimension	ligacao	dia.csv	source	String	Substituir as linhas da source vazias por "Perdido"
dw_Trafego	DimLigacao	dimension	idLigacao				Surrogate Key
dw_Trafego	ftTrafego	factTable	tamanho	dia.csv	tamanho	Float	direto
*O nome do ficheiro csv é referente aos vários dias. Exemplos Segunda, Terça, etc.							

O processo de mapeamento dos dados fonte-destino não foi complicado. Isso deveu-se, em grande parte, à ferramenta de *process mining* que utilizámos. A partir do ambiente da ferramenta, e após termos realizado o processo de mineração, chegou a altura de carregar os dados para uma estrutura, fazer as devidas alterações aos dados e carregar os dados para o *data warehouse*.

casos	source	idTrama	seq	ipOrigem	tamanho	tempo
1	10.0.34.20	27	20417/49487	10.0.34.20	784	2018-05-26 12:40:37
1	10.0.14.1	27	20417/49487	10.0.34.20	784	2018-05-26 13:07:14
1	10.0.24.1	27	20417/49487	10.0.34.20	784	2018-05-26 13:33:51
1	10.0.22.1	27	20417/49487	10.0.34.20	784	2018-05-26 14:00:28
1	10.0.22.2	27	20417/49487	10.0.34.20	784	2018-05-26 14:27:05
2	10.0.28.20	35	32924/40064	10.0.28.20	784	2018-05-26 12:51:55
2	10.0.22.1	35	32924/40064	10.0.28.20	784	2018-05-26 13:29:50
2	10.0.22.2	35	32924/40064	10.0.28.20	784	2018-05-26 14:07:45
3	10.0.40.20	27	2565/1290	10.0.40.20	784	2018-05-26 12:26:16
3	10.0.37.1	27	2565/1290	10.0.40.20	784	2018-05-26 12:38:32
3	10.0.24.1	27	2565/1290	10.0.40.20	784	2018-05-26 12:50:48
3	10.0.22.1	27	2565/1290	10.0.40.20	784	2018-05-26 13:03:04
3	10.0.22.2	27	2565/1290	10.0.40.20	784	2018-05-26 13:15:20
4	10.0.40.20	27	2565/1290	10.0.40.20	784	2018-05-26 13:22:25
4	10.0.37.1	27	2565/1290	10.0.40.20	784	2018-05-26 14:30:50
4	10.0.24.1	27	2565/1290	10.0.40.20	784	2018-05-26 15:39:15
4	10.0.22.1	27	2565/1290	10.0.40.20	784	2018-05-26 16:47:40
4	10.0.22.2	27	2565/1290	10.0.40.20	784	2018-05-26 17:56:05
5	10.0.28.20	35	3266/49676	10.0.28.20	784	2018-05-26 12:25:54
5	10.0.22.1	35	3266/49676	10.0.28.20	784	2018-05-26 12:37:48
5	10.0.22.2	35	3266/49676	10.0.28.20	784	2018-05-26 12:49:42
6	10.0.31.20	27	12015/61230	10.0.31.20	784	2018-05-26 12:34:30
6	10.0.20.1	27	12015/61230	10.0.31.20	784	2018-05-26 12:55:00
6	10.0.22.1	27	12015/61230	10.0.31.20	784	2018-05-26 13:15:30
6	10.0.22.2	27	12015/61230	10.0.31.20	784	2018-05-26 13:36:00
7	10.0.28.20	35	50035/29635	10.0.28.20	784	2018-05-26 13:06:56
7	10.0.22.1	35	50035/29635	10.0.28.20	784	2018-05-26 13:59:52
7	10.0.22.2	35	50035/29635	10.0.28.20	784	2018-05-26 14:52:48

Figura 16- Exemplo dos metadados que serão exportados.

A Figura 16 apresenta uma vista dos metadados que serão exportados para o *data warehouse*. O ficheiro apresenta sete colunas, nomeadamente: casos, *source*, *idTrama, seq, ipOrigem,* tamanho e tempo. As colunas casos, *source,* tamanho e tempo são fundamentais para a análise da qualidade da rede. As restantes colunas foram descartadas. A coluna casos representa por ordem crescente, cada pacote que circula na rede, desde a sua origem até ao seu destino, caso isso aconteça. Na coluna *source* é apresentado todos os pontos por onde o pacote circulou na rede. A coluna tamanho, é o tamanho que esse pacote ocupa na rede. Como os pacotes são todos os mesmos é normal que o tamanho seja todo igual. A coluna tempo é o instante de tempo referente ao momento em que o pacote passou naquela ligação. Posto isto, e comparando com os elementos apresentados na Tabela 2, é fácil fazer o mapeamento entre a fonte e o destino. Para que seja possível realizar operações sobre quais os caminhos percorridos pelos pacotes na rede, os locais por onde os pacotes passaram e o momento em que passaram é necessário que as colunas, casos, *source*, e tempo sejam identificadas como dimensões no *data mart* para que seja possível realizar estas e outras operações sobre os dados.

Como vimos, a partir dos dados fornecidos pela ferramenta de mineração de processos é bastante simples conseguir obter informação bastante pertinente sobre o funcionamento da rede que utilizámos como nosso objeto de trabalho. De referir, por exemplo, que conseguimos saber:

- Quantos pacotes são perdidos na rede por dia?
- Qual o dia da semana no qual circula maior tráfego na rede?
- Quais os locais na rede que apresentam maior tráfego?
- Quais as horas de maior tráfego na rede?
- Quais os dias de maior tráfego na rede?

Perguntas como estas são a base de trabalho de qualquer gestor de um sistema de rede. A sua resposta ajuda dos seus gestores de rede a saberem o que fazer para que se possa melhorar a qualidade do serviço da rede e aumentar a satisfação da experiência de utilização dos seus utilizadores.

4.2 A Estrutura de Dados de Suporte

Para acolhermos a informação que definimos como pertinente para os nossos processos de análise de dados decidimos pela implementação de uma estrutura multidimensional de dados. Pela sua própria natureza, este tipo de estruturas permite-nos facilmente acolher dados temporais e organizá-los de acordo com as várias perspetivas de análise que queremos colocar em prática. O processo de desenvolvimento desta estrutura seguiu o modelo proposto por Kimbal et al. [6] para a conceção e implementação do *data warehouse*. Na sua proposta, Kimbal et al. sugerem-nos uma abordagem mais simples – o método dos "4 passos" - para realizarmos um desenvolvimento mais expedito e minimamente sustentado da estrutura que pretendemos implementar. Na realidade, esta estrutura multidimensional de dados é simplesmente um (pequeno) *data mart* para o acolhimento dos dados relativos ao tráfego do nosso sistema de rede. O método dos "4 passos" define que devemos desenvolver um *data mart* realizando as seguintes operações:

1. Identificar o modelo de negócio.
2. Definir o grão.
3. Identificar as dimensões.
4. Identificar os factos**.**

Tabela 3- A matriz de decisão do sistema.

Caraterização de *Data Mart Trafego*	
Identificação: Tráfego	
Descrição Geral: Informação de suporte para a tomada de decisão sobre o funcionamento da rede, providenciando elementos sobre os percursos dos vários pacotes que circulam na rede.	
Estrutura Base	
Tabela de Factos	*ftTrafego*
Dimensões	
DimData	**X**
DimHora	**X**
DimCasos	**X**
DimLigacao	**X**
Número de Dimensões	4
Tipo	Transacional
Periodicidade	Diária
Descrição	Tráfego gerado pela rede
Utilidade Estratégica	Análise do tráfego da rede. Melhorar a rede disponível.
Utilizadores	Gestores de Rede
Observações	Nada a assinalar

Assim, começamos por definir a nossa matriz de decisão relativa ao *data mart- "Trafego"* - que queríamos implementar (Tabela 3). No nosso caso a matriz é baste óbvia, uma vez que apenas temos um *data mart* e, como, tal todas as dimensões estão apenas relacionadas, em exclusivo com ele. Terminada a matriz de decisão passámos à descrição de cada uma das dimensões identificadas e que integram a sua estrutura base do *data mart.*

O nosso *data mart* apresenta quatro dimensões, sendo elas as seguintes:

1. Dimensão Ligação(*DimLigacao*) (Tabela 4) - IP das várias ligações presentes na rede por onde circulam os vários pacotes. Caso o pacote se tenha perdido o IP aparece "Perdido".

Tabela 4- Caracterização da dimensão *DimLigacao* do *data mart trafego.*

Caracterização da dimensão					
Identificação			*DimLigacao*		
Descrição			Representa todos os endereços de IP por onde circulam os vários pacotes, desde os endereços iniciais até chegar ao destino.		
Tipo			Sem Variação		
Crescimento			0.10% dia.		
Atributos					
Nr	**Identificação**	**Variação [Sim/Não]**	**Domínio**	**Descrição**	**Exemplo**
1	idLigacao	N	Inteiro	Representa o identificador de uma ligação.	1
2	Nome	N	String	Representa o nome da ligação	10.0.0.20
Hierarquia(Ramos)					
Nr	**Identificação**	**Esquema**			
1	H1	idLigacao->Ligacao->All			
Perfis de Utilização					
Gestores de rede e administradores.					
Observações					
Nada a acrescentar					

2. Dimensão Casos(*DimCasos*) (Tabela 5) - número de pacotes diferentes que circulam na rede por dia. Caso haja casos repetidos é preciso proceder a sua substituição.

Tabela 5- Caracterização da dimensão *DimCasos* do *data mart trafego.*

Caracterização da dimensão					
Identificação			*DimCasos*		
Descrição			Representa o número de pacotes diferentes que circulam na rede		
Tipo			Sem Variação		
Crescimento			0.10% dia		
Atributos					
Nr	**Identificação**	**Variação [Sim/Não]**	**Domínio**	**Descrição**	**Exemplo**
1	idCasos	N	Inteiro	Representa o identificador do número de pacotes que circulam na rede	1
2	Casos	N	Inteiro	Representa o número de pacotes que circulam na rede	1
Hierarquia(Ramos)					
Nr	**Identificação**	**Esquema**			
1	H1	idCasos->Casos->All			

Perfis de Utilização
Gestores de rede e administradores.
Observações
Nada a acrescentar

3. Dimensão Hora (*DimHora*) (Tabela 6) - a hora em que um determinado pacote passou numa determinada ligação.

Tabela 6- Caracterização da dimensão *DimHora* do *data mart trafego.*

Caracterização da dimensão					
Identificação			*DimHora*		
Descrição			Horas presentes num relógio		
Tipo			Sem Variação		
Crescimento			Não cresce. É carregado no início do *Data Warehouse* .		
Atributos					
Nr	**Identificação**	**Variação [Sim/Não]**	**Domínio**	**Descrição**	**Exemplo**
1	idHora	N	Inteiro	Identificador na hora	1
2	Hora	N	Inteiro	Hora marcada num relógio	14
Hierarquia(Ramos)					
Nr	**Identificação**	**Esquema**			
1	H1	idHora->Hora->All			
Perfis de Utilização					
Gestores de rede e administradores.					
Observações					
Nada a acrescentar					

4. Dimensão Data(*DimData*) (Tabela 7) - data ao qual os pacotes circularam na rede.

Tabela 7- Caracterização da dimensão *DimData* do *data mart trafego*.

Caracterização da dimensão	
Identificação	*DimData*
Descrição	Calendário do ano e os seus atributos
Tipo	Sem variação
Crescimento	Não cresce. O povoamento desta dimensão é feito durante a fase de arranque do *Data Warehouse* para um período de 2 anos, desde a data mais antiga, i.e.,

			2018-01-01.		
Atributos					
Nr	**Identificação**	**Variação [Sim/Não]**	**Domínio**	**Descrição**	**Exemplo**
1	idData	N	Inteiro	Número único que identifica uma determinada data	1
2	Data	N	Data	Data do calendário	2018-08-30
3	Dia	N	Inteiro	Número do dia do Mês	30
3	Mês	N	Inteiro	Número do Mês	8
4	Semana	N	String	Nome do dia da Semana do ano	Segunda
5	Trimestre	N	Inteiro	Semestre em que o mês se refere	3
7	Ano	N	Inteiro	Ano da data	2018
Hierarquia(Ramos)					
Nr	**Identificação**	**Esquema**			
1	H1	idData->data->mês->trimestre->ano->All			
2	H2	idaData->data->dia->All			
3	H3	idData->data->diaSemana->All			
Perfis de Utilização					
Gestores de rede e administradores.					
Observações					
Nada a acrescentar					

Na etapa relativa à definição do gão a incorporar no *data mart* define-se ao nível de detalhe da informação armazenada no sistema. Posto isto, chegou a altura de escolher o grão, isto é, a peça de dados mais elementar que é possível obter do modelo dimensional. O grão escolhido foi o tráfego recolhido numa dada ligação para um determinado dia.

Apresentadas e descritas as diversas dimensões de análise, podemos nesta altura resumir a sua definição através da Tabela 8.

Tabela 8- Apresentação e descrição das dimensões de análise.

Dimensões *Data Mart* Tráfego			
Nr	Identificação	Descrição	Esquema(Tipo)
1	Casos	Identifica o caminho percorrido pelos vários pacotes	*DimCasos* (Sem variação)
2	Ligação	IP da ligação por onde cada pacote passou	DimLigacao (Sem variação)
3	Hora	Dimensão temporal. Acolhe as horas do tráfego ocorrido na rede	DimHora (Sem variação)
4	Data	Dimensão temporal. Acolhe todos os atributos ao longo do tempo, como data, dia da semana, mês, trimestre e ano	DimData (Sem variação)

Como forma de resumir o processo de modelação dimensional, veja-se o esquema dimensional produzido para o *Data Mart Trafego* (Figura 17), realizado com base na notação *Golfarelli et al.* [5], usando a ferramenta *draw.io* [18].

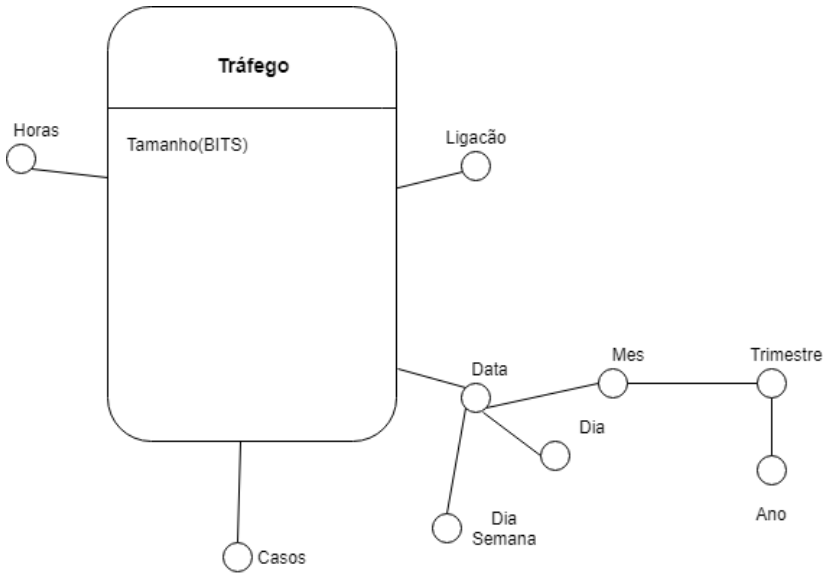


Figura 17-Esquema estrela para o acolhimento dos registos de tráfego da rede.

Todos os processos de negócios são representados por um modelo dimensional que consiste numa tabela de factos que irá conter os eventos de medições numéricas, envolvendo um conjunto de tabelas de dimensão que irá conter o contexto textual, assumindo como verdadeiro quando foi carregado. O primeiro objetivo da tabela de factos é que seja simples e que armazena toda a informação relevante para a caracterização dos seus factos. Desta forma, os utilizadores beneficiam dessa simplicidade, uma vez que os seus dados serão fáceis de entender e de navegar.

Tabela 9- Caracterização da tabela de factos *"Ft_Trafego"*.

Caracterização da tabela de factos					
Identificação			*Ft_Trafego*		
Descrição			Tabela que acolhe o tráfego da rede.		
Data *mart*			Comercial		
Tipo			Transacional		
Utilidade estratégica			Melhorar a internet disponibilizada aos alunos. Acompanhamento do tráfego gerado pelos alunos.		
Povoamento			Realizado diariamente entre as nove horas e onze horas da noite.		
Dimensão inicial					
Crescimento			0.10% dia.		
Período de dados			Desde o ano de 2018. Os anos anteriores ficarão em arquivos.		
Atributos					
Dimensões					
Nr	**Identificação**	**Chave**	**Domínio**	**Descrição**	**Exemplo**
1	IdLigacao	S	Inteiro	Código interno do nome da ligação	1
2	IdCaso	S	Inteiro	Código interno do caso referente a um pacote	1
3	IdData	S	Inteiro	Código da data referente a data em que o pacote circulou na rede.	1
3	idHora	S	Inteiro	Código da hora referente a hora em que o pacote circulou na rede	1
Medidas					
Nr	**Identificação**	**Domínio**	**Descrição**	**Exemplos**	
1	Tamanho	Float	Tamanho de cada pacote na rede.	2	
Índice					
Nr	Identificação	Tipo	Descrição		
1	IdTrafego	Primário	Único, ordenado fisicamente(*clustered*) de forma crescente.		
2	IdCaso	Secundário	Ordenado de forma crescente.		

3	IdData	Secundário	Ordenado de forma crescente.
4	idHora	Secundário	Ordenado de forma crescente.
5	idLigacao	Secundário	Ordenado de forma crescente.
Perfis de Utilização			
Administrador da base de dados e gestores de rede			
Observações			
Todos os valores considerados nos atributos medida são em bits. Qualquer valor que não esteja em bits, deve primeiramente ser convertida em bits.			

A tabela de factos definida ("*Ft_Trafego*") (Tabela 9) acolhe os dados relativos ao tráfego que circulou na rede. Esta tabela permite conhecer o caminho por onde os pacotes passam, a hora em que eles passam e o dia em que eles passam. Por sua vez, o tamanho de cada pacote irá permitir perceber a "carga" que circula na rede. A tabela de factos e as tabelas de dimensão com as quais está relacionada podem ser vistas na Figura 18, que apresenta o esquema lógico que define a estrutura do *Data Mart "Trafego"*.

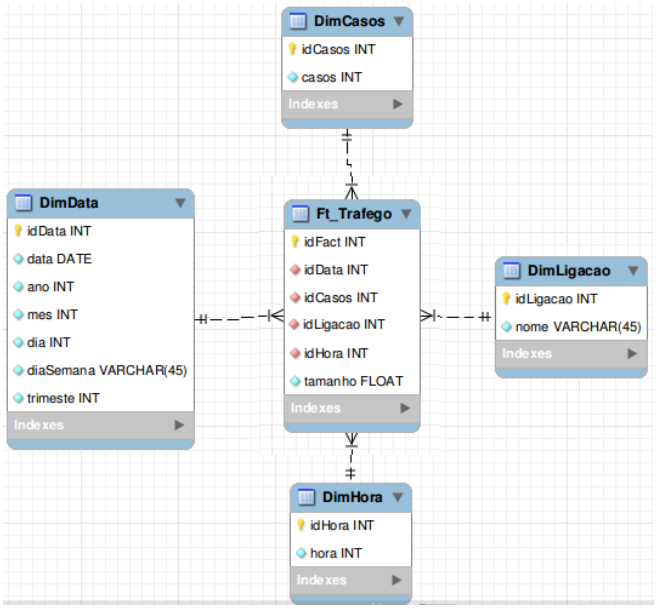


Figura 18- Esquema lógico para o *Data Mart* Tráfego.

O esquema lógico apresentado na Figura 18 é composto pelas dimensões *"DimLigacao"*, *"DimCasos"*, *"DimData"* e *"DimHora"*, em que ao centro está a tabela de factos *"Ft_Trafego"*. O identificador único da tabela *"Ft_Trafego"* é o único que é auto numerado, sendo que os restantes identificadores serão atribuídos no processo de carregamento do ETL para o *data mart* idealizado. A tabela de factos apresenta apenas uma única medida, onde podemos aplicar sobre ela algumas funções de agregação, como é o exemplo da soma, média, máximo, entre outras. Quando terminado a construção do esquema lógico passamos à fase seguinte, que é transformar o modelo lógico em modelo físico. Através do modelo físico conseguimos interrogar o *data mart* com as perguntas que achámos mais pertinentes. Exemplos dessas perguntas podem ser encontrados nas Figuras 19 e 20.

```
select mes,nome, count(*) as 'Número de pacotes por mês em cada ligação'
from FactTable AS Ft, DimData as DD,DimLigacao AS DL
where Ft.idData= DD.idData and DL.idLigacao=Ft.idLigacao
group by mes, nome
```

Figura 19- Número de pacotes que circula em cada mês em cada ligação.

Na Figura 19 interrogámos o *data mart* sobre o número de pacotes por mês que circula em cada ligação IP. Esta *query* devolve a informação agrupada por mês e pelo IP de cada ligação.

```
select hora,nome, count(*) as 'Número de pacotes que circulam por hora em ligação'
from FactTable AS Ft, DimHora as DH,DimLigacao AS DL
where Ft.idHora= DH.idHora and DL.idLigacao=Ft.idLigacao
group by hora, nome
```

Figura 20- Número de pacotes que circula por hora em cada ligação.

Na Figura 20 o interrogámos o *data mart* sobre o número de pacotes por hora que circula em cada ligação IP. Esta *query* devolve a informação agrupada por hora e pelo IP de cada ligação.

4.3 O Povoamento da Estrutura Multidimensional

Após a definição do modelo dimensional do *data warehouse*, é necessário proceder à definição e caracterização do processo que o irá povoar. Inicialmente, é necessário proceder à identificação dos diferentes carregamentos, que neste caso irão ser dois, o carregamento inicial e regular do *data warehouse.*

O carregamento inicial é realizado apenas uma vez, no arranque do sistema, que irá permitir efetuar um povoamento inicial das tabelas de dimensão e da tabela de factos do *data warehouse*. Posteriormente, é necessário definir um processo ETL que represente o carregamento regular do sistema, que irá ocorrer diariamente segundo uma janela de oportunidade. Este carregamento caracteriza-se por atualizar a informação do *data warehouse* com as novas informações.

Em seguida iremos descrever as várias etapas que os dados passaram até serem carregados para o *data warehouse.* Assim organizámos o processo de povoamento nas seguintes etapas:

1. Extração, processo ao qual carregamos a informação retirada do ficheiro *log* para uma base de dados *mysql* [34] para ser analisada, processada e depois carregada para o *data warehouse.*
2. Limpeza, etapa ao qual substituímos os valores do IP que estavam a nulo por "Perdido" e eliminámos os atributos irrelevantes.
3. Conformidade, uma das etapas mais trabalhosas, em ordem a garantir integridade ao *data warehouse* foi necessário garantir que não era inserido duas vezes o mesmo endereço de IP nas tabelas de *Surrogate Key*, visto que são estas tabelas responsáveis pela integridade das dimensões presentes no *data warehouse*. O mesmo acontece para os *casos* presentes nos ficheiros. Dias diferentes apresentam *casos* iguais, mas são pacotes diferentes, por isso é necessário garantir que cada pacote tenha um identificador único, o que irá originar uma substituição do identificador do pacote por um outro que ainda não esteja a ser utilizado.
4. Conciliação, esta etapa tem apenas como função inserir os registos em tabelas temporárias, para depois serem carregado para o *data warehouse.*
5. Carregamento, consiste em carregar o conteúdo das tabelas temporárias para o *data warehouse.* O conteúdo presente nas tabelas temporárias uma vez transferido para o *data warehouse* é apagado das tabelas temporárias.

Na realidade, o processo de povoamento que desenvolvemos é bastante convencional, uma vez que não tem qualquer etapa menos convencional. Além disso, a natureza e os dados da fonte de dados que tivemos à nossa disposição facilitaram também bastante a conceção dos processos bem como a sua posterior execução. Para tornar a explicação do processo de povoamento mais clara e

compreensível, optámos por utilizar modelos *BPMN* [7,8] para fazermos a modelação de cada uma das etapas de povoamento e sustentar a sua explicação de forma gráfica, complementada por descrições simples e claras. Para desenvolvermos os modelos *BPMN* utilizámos a ferramenta *Visual Paradigm* [9] e para fazermos a sua implementação usamos a ferramenta *Pentaho Data Integration* [14]. Vejamos, então, cada uma das etapas referidas em particular.

4.3.1 Extração dos dados

O modelo da etapa de extração de dados será apresentado na Figura 21. Nele podemos ver que a primeira tarefa a ser realizada é o carregamento de ficheiro *CSV* para uma base de dados *mysql*.

Figura 21-Extração dos dados de tráfego realizado na segunda-feira.

Depois de recolhida o tráfego realizado na Segunda-feira, extraímos essa informação para uma tabela presente no ETL. Essa tabela irá conter todos os registos do tráfego para mais tarde ser tratado e depois carregado para o *Data Warehouse.*

No entanto, as dimensões Hora (Figura 22) e Data (Figura 23), são excecionalmente carregadas de imediato para o *data mart*, isto é, no carregamento inicial do processo de ETL, pois são dimensões sem variação e que se sabe à partida que não terão qualquer tipo de falhas, visto que foram geradas automaticamente sob a forma de um ficheiro *CSV*.

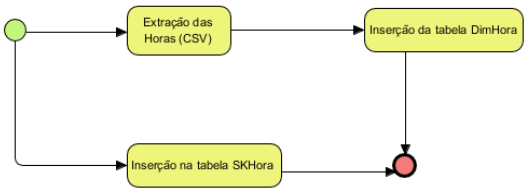


Figura 22- Processo de extração e carregamento da dimensão Hora no *data mart*.

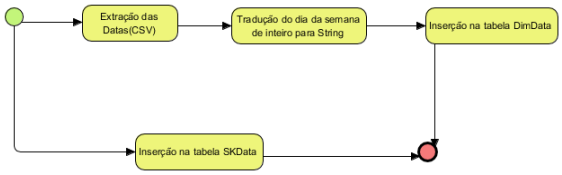


Figura 23-Processo de extração e carregamento da dimensão Data no *data mart*.

4.3.2 Limpeza dos dados

A segunda etapa do processo de povoamento é a limpeza dos dados angariados. Nesta etapa (Figura 24) procedemos à eliminação atributos que tinham relevância para o caso de estudo em questão, em particular os atributos "*seq*", "*IpOrigem*" e "*idTrama*". Na realidade, o que fizemos foi fazer uma seleção da informação que era importante para o modelo de análise que queríamos implementar. Depois dessa tarefa ter sido executada, passámos à tarefa de limpeza dos dados propriamente dita. Na tarefa de limpeza carregamos os registos selecionados para as tabelas de limpeza e atualizamos os endereços de IP a nulo por "Perdido".

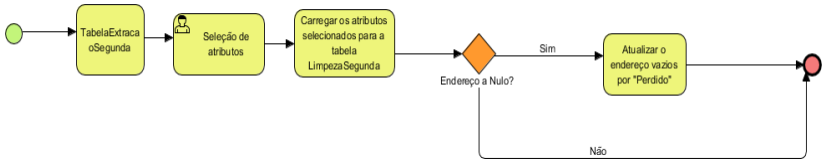


Figura 24- Limpeza do tráfego realizado na segunda-feira.

4.3.3 Conformidade

A terceira etapa no processo de povoamento é a conformidade dos dados. Nesta etapa (Figura 25) os dados começam a ser preparados para mais tarde serem carregado para o *data mart* idealizado. Para tal é necessário eliminar a redundância nos dados. Como são efetuas várias recolhas de tráfego em diferentes dias e as recolhas são independentes umas das outras a coluna *casos*, isto é, o número de pacotes que circula na rede num determinado dia, irá conter números repetidos nos vários dias. Sendo pacotes diferentes não podem aparecer com o mesmo número, isso leva a uma

redundância dos dados e consequentemente conclusões erradas. Para evitar que isso aconteça é necessário proceder à substituição do número de casos, por um outro que ainda não tenha sido utilizado (Figura 26). No caso dos endereços de IP é necessário verificar se os endereços de IP já foram inseridos para evitar inserir endereços de IP repetidos (Figura 27).

Figura 25- Processo de Conformidade.

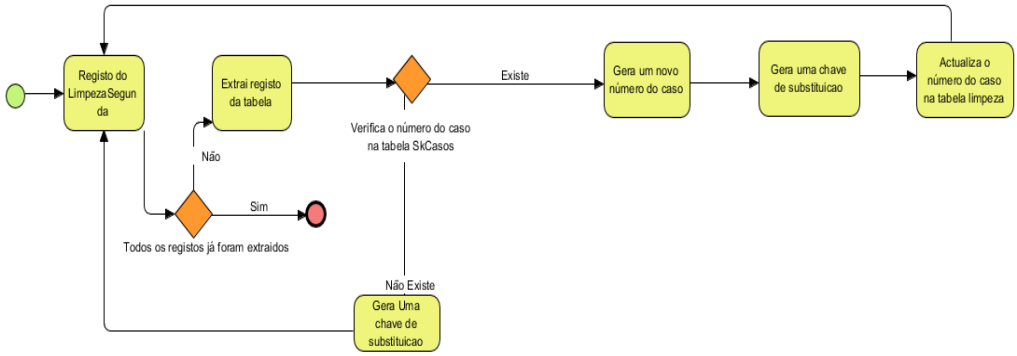


Figura 26- Subprocesso de geração de chaves para os casos.

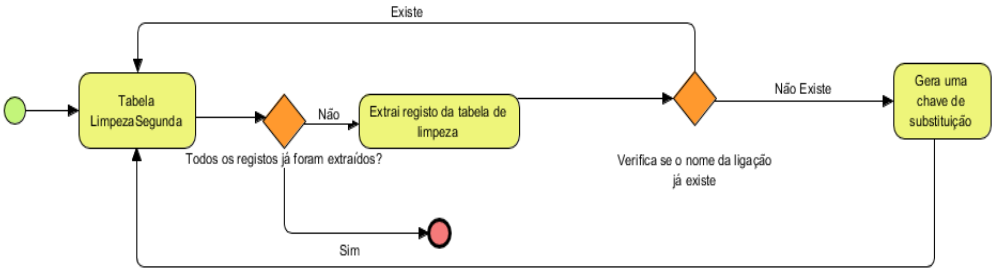


Figura 27- Subprocesso de geração de chaves para os endereços de IP.

4.3.4 Conciliação

A última etapa antes do carregamento dos dados para o *data mart* é a conciliação (Figura 28). Nesta etapa os dados são inseridos em tabelas temporárias, "*tmpDimLigacao*", "*tmpDimCasos*" e "*tmpFt_Trafego*" para depois serem carregados para o *data mart,* para as respetivas tabelas de dimensões e de factos.

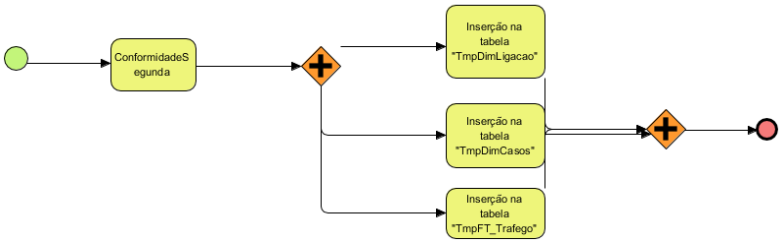


Figura 28- Carregamento dos dados em tabelas temporárias.

4.3.5 Carregamento dos Dados

Por fim chegou a altura de carregar os dados para o *data mart* (Figura 29). Este carregamento é bastante simples e traduz-se na passagem dos dados presentes em tabelas temporárias para o *data mart.* Neste carregamento é preciso carregar, primeiramente, os dados presentes nas tabelas temporárias de dimensões para as tabelas de dimensões do *data mart trafego* e só depois é que é carregado os dados presentes na tabela de factos temporária para a tabela de factos presente no *data mart trafego.*

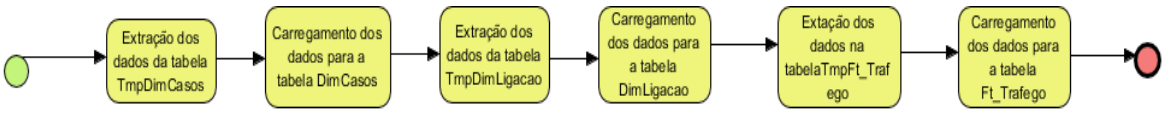


Figura 29- Carregamento para o *data mart trafego.*

4.4 A implementação do Processo de Povoamento

Tendo todos os elementos do sistema de povoamento bem identificados e descritos, tratamos de fazer a implementação do processo. Para isso utilizámos a ferramenta *Pentaho Data Integration*

(Kettle) [14]. Como referido anteriormente existem dois tipos de povoamentos, o povoamento inicial e o povoamento regular.

A Figura 30 representa uma vista geral sobre as várias etapas de desenvolvimento no carregamento inicial para o *data mart*. Quando o carregamento estiver concluído o *Kettle* irá apagar toda a informação presente nas tabelas temporárias para evitar que as mesmas sejam novamente carregadas para o *data mart.*

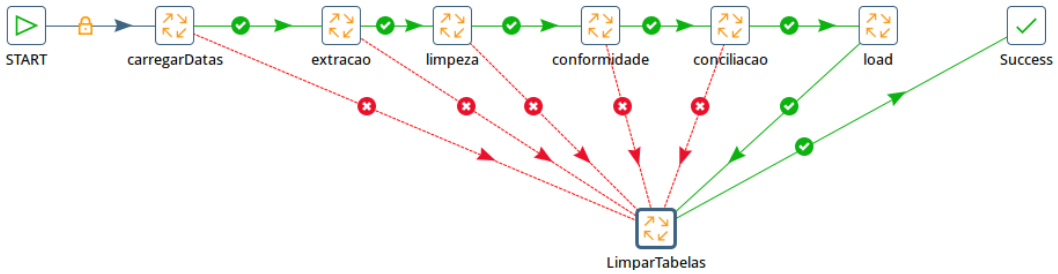


Figura 30- Processo ETL desenvolvido para o povoamento inicial do *data mart trafego.*

A Figura 31 ilustra as várias etapas do povoamento regular do *data mart*. Este povoamento é muito parecido ao povoamento inicial, só que desta vez já não é preciso carregar as datas e as horas, uma vez que as mesmas já se encontram no *data mart*. Por fim e após a limpeza das tabelas temporárias do ETL é necessário atualizar a cache da base de dados multidimensional, para que seja possível carregar para a base de dados multidimensional os novos registos.

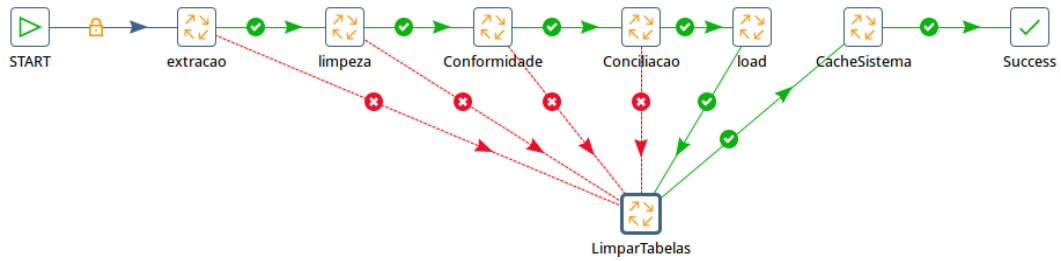


Figura 31- Processo ETL desenvolvido para o povoamento regular do *data mart trafego.*

A Figura 32 representa o processo de inserção da data e da hora nas respetivas tabelas de dimensões, assim como nas respetivas tabelas de *surrogate key*.

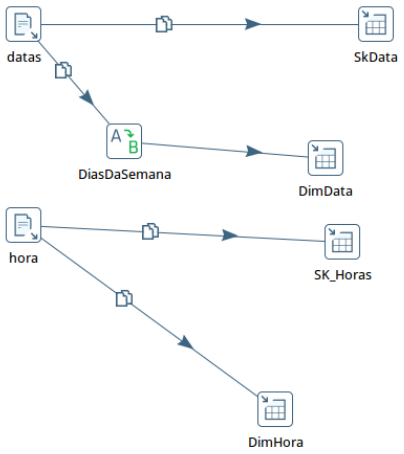


Figura 32- Inserção da data e da hora no *data mart trafego.*

A Figura 33 caracteriza um exemplo da extração dos dados capturados a uma segunda-feira do ficheiro *CSV* para a base de dados *mysql*, nomeadamente para a tabela *ExtracaoSegunda*.

Figura 33- Extração dos dados do tráfego rede capturados na segunda-feira.

A Figura 34 apresenta uma descrição geral sobre o processo de limpeza. Primeiramente começamos por selecionar os atributos relevantes para o processo de análise. Uma vez esses atributos serem selecionados, são carregados para as tabelas de limpeza. Uma vez esses atributos carregados atualizamos os valores dos endereços de IP a nulo para "Perdido".

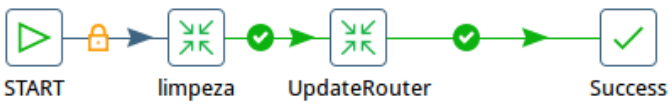


Figura 34- Processo geral da limpeza dos dados.

A Figura 35 ilustra o processo de conformidade dos dados. Começamos por gerar as chaves de substituição (*surrogate key*) para as ligações de IP presentes nas tabelas de limpeza. Caso já existe uma chave para aquela ligação, essa é a chave utilizada. Depois geramos chaves para os *casos* presentes na rede. Caso haja *casos* repetidos procedemos à substituição desses *casos* por outros, atualizando a informação nas respetivas tabelas de limpeza. Quando este trabalho estiver concluído, os dados serão carregados para as tabelas de conformidade.

Figura 35- Processo geral de conformidade dos dados.

A Figura 36 apresenta o processo de conciliação, que não é nada mais do que carregar os registos presentes nas tabelas de conformidade e carrega-los para as tabelas temporárias de dimensões e na tabela temporária de factos. Para não haver problemas de integridade referencial, isto é, uma chave estrangeira não ter associada uma chave primária na tabela que faz referência, é necessário, primeiramente, carregar os valores presentes nas tabelas temporárias de dimensão. Para carregar a informação para a tabela de factos é necessário recorrer a um *stored procedure*. Este *stored procedure* (Figura 37) certifica-se que os dados inseridos na tabela de factos apresentam os mesmos identificadores dos registos inseridos nas tabelas de dimensões temporárias. O *stored procedure* recebe um conjunto de atributos presentes na tabela de conformidade, os *casos*, o IP da ligação e o instante que o pacote passou naquela ligação, vai as tabelas de *surrogate key* e retira o identificador referente aos vários registos. Uma vez tendo esses identificadores inseri-os na tabela de factos temporária.

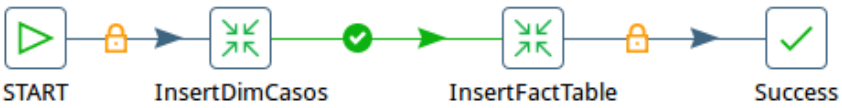


Figura 36- Processo geral de conciliação dos dados.

```
DELIMITER $$
create procedure insertFactTable(in in_casos INT, in in_router VARCHAR(45)
    , in in_time timestamp, in in_tamanhoPacote float)
begin
declare idCasosTmp int;
declare idRouterTmp int;
declare idDataTmp int;
declare idHoraTmp INT;

select idCasos into idCasosTmp
from SkCasos
where numeroCasos=in_casos;

select idRouter into idRouterTmp
from SkRouter
where nome=in_router;

select idData into idDataTmp
from SkData
where data=date(in_time);

select idHora into idHoraTmp
from SkHora
where hora =hour(in_time);


insert into factTable (idRouter,idData,idCasos,idHora,tamanhoPacote)
    values
    (idRouterTmp,idDataTmp,idCasosTmp,idHoraTmp,in_tamanhoPacote);
end $$
DELIMITER ;
```

Figura 37-*Stored procedure* usado para a inserção dos registos na tabela de factos temporária.

Uma vez que os registos sejam inseridos nas tabelas temporárias chegou a altura de serem carregados para o *data mart*. Por fim é guardado o instante em que os dados foram carregados para o *data mart* (Figura 38).

Figura 38-Carregamento dos registos para o *data mart*.

4.5 Base de Dados Multidimensional

Depois de definida a estrutura do *data mart* passámos à criação da correspondente base de dados multidimensional (CUBO OLAP). Para tal foi utilizada a ferramenta *Pentaho* [10], pelas suas características de integração com a ferramenta utilizada anteriormente na implementação do processo de povoamento e pela sua facilidade de utilização e compreensão. Esta ferramenta utiliza o motor analítico *Mondrian* [11] para suportar as *queries* em *MDX* [12], que se pretendam desenvolver sobre a estrutura multidimensional criada.

```
<Cube name="DWGrupos" visible="true" cache="true" enabled="true">
  <Table name="FactTable">
  </Table>
  <DimensionUsage source="DimData" name="DimData" visible="true" foreignKey="idData" highCardinality="false">
  </DimensionUsage>
  <DimensionUsage source="DimHora" name="DimHora" visible="true" foreignKey="idHora" highCardinality="false">
  </DimensionUsage>
  <DimensionUsage source="DimLigacao" name="DimLigacao" visible="true" foreignKey="idLigacao" highCardinality="false">
  </DimensionUsage>
  <DimensionUsage source="DimCasos" name="DimCasos" visible="true" foreignKey="idCasos" highCardinality="false">
  </DimensionUsage>
  <Measure name="tamanhoPacote" column="tamanhoPacote" datatype="Numeric" formatString="#" aggregator="sum" visible="true">
  </Measure>
</Cube>
```

Figura 39-Fragmento de um ficheiro XML com a definição de estrutura de dados multidimensional.

O processo de tradução da base de dados para o ambiente da ferramenta *Pentaho* não é, porém, direto. Primeiro, é preciso primeiro converter a base de dados num formato *XML* (Figura 39) e, depois, fazer a sua importação. Para realizar essa operação utilizámos o *Pentaho Schema Workbench* [13], que permite a criação e teste de *Mondrian OLAP cube schemas*. Na Figura 40 podemos ver uma imagem do ambiente dessa ferramenta já com a nossa estrutura de dados importada, assim como as suas dimensões e as respetivas hierarquias que foram retiradas do *data mart* implementado.

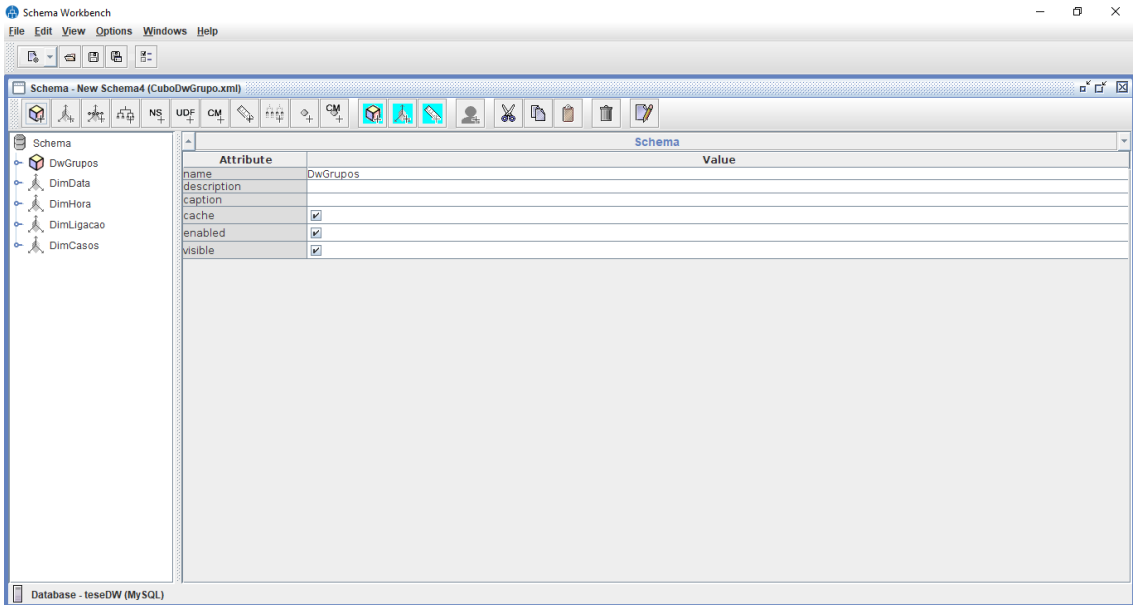


Figura 40- O ambiente da ferramenta *Schema Workbench.*

Em termos gerais, o cubo implementado contém quatro dimensões, nomeadamente: *"DimHora", "DimData", "DimCasos"* e *"DimLigacao"*. Todas estas dimensões são retiradas diretamente do *data mart* já implementado. Em relação as medidas, visto que o *data mart* apresenta apenas uma medida, -" tamanho" -, esta será, pois, a única a figurar no cubo implementado. No que diz respeito as hierarquias, veja-se de seguida aquelas que foram definidas (organizadas por dimensão de análise):

DimData

H1: idData->data->mês->trimestre->ano->All

H2: idaData->data->dia->All

H3: idData->data->diaSemana->All

DimCasos

H1: idCasos->Casos->All

DimLigacao

H1: idLigacao->Ligacao->All

DimHora

H1: idHora->Hora->All

As hierarquias quem compõem a dimensão temporal, "*DimData*" (Tabela 7), permitem agregar ou desagregar a medida pela data em que foram recolhidas, pelo mês, trimestre e por fim por ano. Também é possível agregar pelo dia em que foi feita a recolha dos dados, assim como pelo dia da semana.

Capítulo 5

Um Índice de Monitorização

5.1 Estabelecimento do índice

O desempenho de uma rede é algo que regularmente preocupar os seus administradores e gestores de rede. Ter acessos rápidos e seguros à Internet são aspetos bastante pertinentes e discutidos atualmente. Com o aparecimento dos *smartphones* a Internet passou a estar praticamente em todo lado, sendo utilizada a qualquer hora. Por isso, é importante que uma rede consiga aguentar os acessos de volume de dados que esses e outros dispositivos provocam. Para garantir uma qualidade de serviço da rede adequada a esse tipo de utilização, é necessário que os administradores dos sistemas de rede façam um acompanhamento constante do estado da rede e daquilo que nela se passa para evitar quebras de serviço ou mesmo ruturas.

Na maioria dos casos, os utilizadores gostam de saber o que se passa com a sua rede em que estão integrados, ou seja, saber coisas como, a velocidade da rede, o volume de tráfego que estão a gerar, saber o tempo em que a sua rede demora a estabelecer uma ligação, as suas vulnerabilidades, entre outros fatores. Muitas vezes é informação apenas de curiosidade, mas na realidade são elementos que nos ajudam a utilizar de forma mais efetiva um sistema de rede. Todavia, para se conhecer elementos como esses é necessário ter acesso (e saber usar) algumas ferramentas que permitem fazer a monitorização daquilo que está a acontecer na rede e informar sobre o tráfego que circula na rede [15].

Tendo acesso à informação sobre o tráfego de uma rede consegue-se conceber e implementar processos de análise bastante efetivos sobre a qualidade de serviço da própria rede. Porém, não raras vezes, estes processos de análise não têm em conta alguns fatores bastante pertinentes que afetam a qualidade do serviço. Muitas dessas ferramentas apenas têm em conta o volume de tráfego gerado na rede, o que, em termos gerais, não é suficiente para se retirar colusões sobre a qualidade do serviço. É preciso, também, ter em conta outros fatores, como por exemplo o tempo que um dispositivo demora a estabelecer uma conexão [16], ou seja, o tempo que demora desde que o utilizador faz um pedido à rede, por exemplo aceder a um website, até esse pedido ser aceite, o tempo de processamento de um router, isto é, o tempo em que o router recebe um pacote o analisa e mediante dessa análise decidir o que faz com esse pacote. Fatores como estes afetam também o desempenho da rede. Todavia, estes fatores não se conseguem obter através de uma simples recolha de tráfego, sendo necessário utilizar outro tipo de ferramentas para se obter esse tipo de informações.

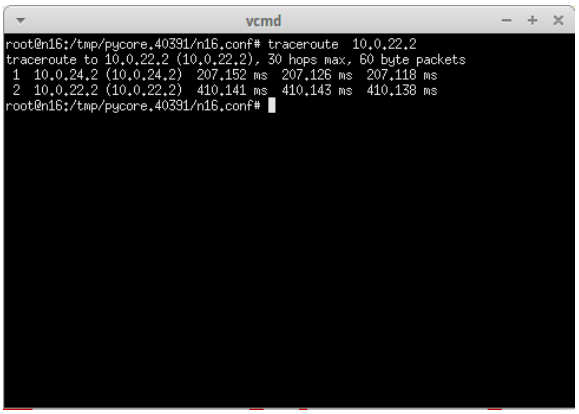


Figura 41-Exemplo do resultado da aplicação do comando *traceroute* sobre um certo dispositivo.

Como vimos anteriormente, com a ajuda da ferramenta *Disco* conseguimos traçar um mapa da rede que idealizámos com base no tráfego que nela foi verificado, bem como retirar outros elementos acerca da rede, como seja o caso dos pacotes enviados por cada router, dos pacotes perdidos e dos vários tempos, como a média, mediana que os pacotes demoram a passar de ligação em ligação, ou então o valor mínimo ou máximo, que os pacotes demoraram a transitar de ligação. Com esta informação, já é possível pensar em algo que nos permita de uma forma

simples, mas efetiva, medir a cada momento a qualidade de serviço da rede em cada ponto de ligação. Porém a ferramenta *Disco* não permite saber o tempo que se demora a estabelecer uma ligação em cada ponto da rede. Assim, tivemos que utilizar um outro comando "*traceroute ipDestino*" [17] para obter esse tempo[7]. Esse comando envia sempre três pacotes ao endereço destino e como tal retorna apenas o tempo que cada pacote demora a chegar a esse destino. Isso fez com que tivéssemos de fazer uma média do tempo que demora a estabelecer uma conexão com base na informação extraída sobre os três pacotes (Figura 41).

Todos os elementos que tivemos o cuidado de angariar com base na monitorização da rede visava o estabelecimento de um índice de monitorização. Os índices deste tipo permitem-nos de uma forma simples ter uma ideia bastante concreta acerca daquilo que estão a medir, desde que, obviamente, saibamos interpretar os valores que ele vai revelando ao longo do tempo. Numa primeira tentativa para estabelecer o referido índice de monitorização do serviço de uma rede, estabelecemos uma primeira expressão para o seu cálculo que apenas determinava a percentagem de pacotes perdidos em cada ligação num determinado dia. A fórmula de cálculo estabelecida nesta primeira fase foi a seguinte:

$$P_{d,r} = \frac{Pacotes\ Perdidos}{Pacotes\ Totais}$$

A fórmula $P_{d,r}$ resulta da informação obtida pela ferramenta *Disco,* onde *d* significa o dia em que a recolha foi feita e o r o router/ligação onde essa recolha foi efetuada.

Todavia, como sabemos, ter apenas a percentagem de pacotes perdidos revela um índice bastante fraco, por informativo e pouco relevante. Para enriquecer esta primeira expressão decidimos adicionar mais alguns elementos de forma a fortalecer a qualidade da fórmula e do próprio índice. Um fator que usualmente tem grande impacto na qualidade do serviço de uma rede é o volume de tráfego que passa numa dada ligação, se a velocidade da ligação aguentar esse volume de tráfego. A velocidade da ligação pode ser obtida perguntando aos administradores da rede. Assim somando o tamanho de todos os pacotes que circulam numa dada ligação, num determinado dia e dividi-los pela velocidade da rede, conseguimos obter o tempo que aquela ligação demora a despachar

[7] Apesar que na pratica este comando não sirva para se obter o tempo de ligação entre um dispositivo e outro, para enaltecer a fórmula do índice optou-se por usar tempo devolvido por este comando. Na prática era necessário recorrer a uma ferramenta para se obter o tempo de ligação entre um dispositivo e outro.

todos os pacotes que circulam na rede. A expressão seguinte revela a fórmula como esse valor é calculado:

$$V_{d,\,r} = \frac{\sum tamanho\ Pacotes}{Velocidade\ Rede}$$

Outros aspetos pertinentes também a ter em conta é o tempo que uma ligação demora a estabelecer uma conexão($T_{d,r}$) e o tempo que uma dada ligação demora a processar um pacote ($TP_{d,r}$), ou seja, o tempo em que a ligação demora a receber o pacote, a analisá-lo e a decidir o que irá fazer com ele, este tempo foi obtido usando a ferramenta *Disco*.

Para obtermos uma ideia bastante clara da qualidade revelada pelo índice definimos que este iria trabalhar numa escala de 0 a 1, na qual o valor 0 significa boa qualidade da rede e o 1 significa péssima qualidade da rede. Para isso, foi necessário desenvolver uma fórmula que desenvolvesse o valor do índice nesse intervalo de valores. Primeiramente, foi necessário estabelecer um peso para cada elemento (variável) da fórmula, uma vez que nem todos os elementos influenciam a qualidade do serviço da mesma maneira. Por isso optou-se por atribuir diferentes pesos a cada uma das fórmulas. Como o número de pacotes perdidos apresenta algum impacto na qualidade das ligações optámos por atribuir à percentagem de pacotes perdidos um peso de 20%. Porém, se a velocidade da rede for grande, rapidamente a rede consegue recuperar qualquer pacote perdido. Logo não é, pois, maior fator na qualidade da ligação.

Os elementos com maior influência sobre a qualidade do serviço da rede são o tempo que cada ligação demora a despachar todos os pacotes que circulam na rede, bem como o tempo de processamento de cada ligação. Como tal, a cada um destes elementos foi atribuído um peso de 35% na qualidade das ligações. Como estas duas fórmulas retornam valores muito elevados foi necessário recorrer a uma função que devolvesse o valor desse elemento numa escala de valores entre 0 e 1, em que 0 é o melhor valor e 1 o pior valor possível. Tudo isto para que os diversos elementos se enquadrassem na mesma escala de valores do índice de qualidade. A função definida para essa padronização de valores foi a seguinte:

$$F(x) = \lim_{x \to +\infty} \frac{X}{x+1} = 1$$

Assim, quanto maior for o valor retornado pelas funções de cada um dos elementos que entram no cálculo do índice pior será, obviamente, o índice das várias ligações do sistema de rede em análise.

Por fim, temos o tempo de conexão, que apresenta algum impacto na qualidade das ligações. Porém, o impacto desse elemento por vezes não chega a ser notado[8]. Devido a isso atribuímos-lhe uma importante menor, com um peso menor (10%) no cálculo do índice. Tendo, assim, todos os elementos com influência no cálculo do nosso índice de qualidade de serviço, definimos, por fim, a seguinte fórmula para o estabelecimento do índice referido:

$$\text{Índice}_{\text{bem-estar d, r}} = 0.2*P_{d,r} + 0.35*f(V_{d,r}) + 0.1*T_{d,r} + 0.35*f(TP_{d,r})$$

Nesta fórmula, tal como referido, podemos ver os vários elementos no cálculo do índice de bem-estar da rede, com a influência de cada um dos seus correspondentes pesos. No desenvolvimento da fórmula foi tido em conta vários fatores que pudessem enriquecer o índice de bem-estar. Porém a maneira como estes fatores influenciam a qualidade da rede é difícil de medir e varia de gestor para gestor. Também é preciso ter em conta outros fatores, como por exemplo, por muito rápido que seja a velocidade da rede se houver muito congestionamento na rede, os pacotes perdidos acabam por ter um grande impacto na qualidade do serviço, visto que será necessário voltar a enviar novamente estes pacotes o que irá provocar um maior congestionamento. Como tal é impossível prever o real peso que estes elementos apresentam na qualidade da rede, ficando assim ao critério de cada um, depois de várias análises feitas à rede, atribuir os pesos.

5.2 O *Dataset* usado no cálculo do índice

Para calcular o índice da qualidade de serviço não podíamos usar o mesmo *dataset* (Figura 42) que foi usado anteriormente, uma vez que os elementos de dados requeridos são obviamente, diferentes. Assim, preparámos um segundo *dataset*, especialmente organizado para o suporte do

[8] Neste trabalho só estamos a ter em conta o tempo que ele demora a conectar-se a um determinado site, isto na prática nunca demorará mais que uns milissegundos e o cliente por vezes nunca chega a dar conta da troca do DNS. É claro que trocar de DNS, ou seja, optar por um outro que não aquele fornecido pelos distribuidores de internet, um exemplo claro é o DNS da Google, levanta uma série de questões que não serão abordadas nesta tese.

cálculo do índice. Na Tabela 10 estão apresentados os vários atributos que constituem esse *dataset*, bem como a sua respetiva descrição.

Tabela 10 Descrição dos vários atributos usados no cálculo do índice.

NomeLigação	Apresenta as várias ligações intermédias existentes na rede. Não contém as ligações inicias visto que estás apenas geram o tráfego, logo não são o problema da rede.
TotalPacotes	Número de pacotes que circulam naquela ligação
TotalPerdidos	Número de pacotes perdidos naquela ligação
PercentagemPerdidos	Razão entre o número de pacotes perdidos com o número de pacotes totais
TamanhoTotal	Tamanho dos vários pacotes que circulam naquela ligação
Velocidade	Velocidade daquela ligação
TempoTransporte	Tempo que a ligação demora a transportar os pacotes que por lá circulam
Data	Data da recolha do tráfego
TempoNomes	Tempo que uma dada ligação demora a estabelecer a conexão com o endereço de destino
TempoProcessamento	Tempo que cada ligação demora a processar os vários pacotes que nela circulam
ValorIndice	A qualidade daquela ligação

NomeLig	TotalPacc	TotalPerd	Percenta(	TamanhoTo	Velocidad	TempoTransp	Data	TempoNo	TempoProce	ValorIndice
10.0.3.1	107815	168	0.001558	85968736	512000	167.907688	2018-May-27	0.609071	1588.039	0.735872
10.0.6.1	218277	585	0.00268	164207232	512000	320.71725	2018-May-27	0.506492	1606.4245	0.732851
10.0.10.1	329173	423	0.001285	235506544	512000	459.973719	2018-May-27	0.405588	1692.26075	0.728146
10.0.17.1	328729	742	0.002257	233888368	512000	456.813219	2018-May-27	0.302724	1566.12575	0.722701
10.0.20.1	656467	1541	0.002347	451508096	512000	881.85175	2018-May-27	0.202979	1639.46225	0.716733
10.0.22.1	1441921	209978	0.145624	947305860	512000	1850.206758	2018-May-27	0.102167	1755.44922	0.738006
10.0.14.1	113431	196	0.001728	88150608	512000	172.169156	2018-May-27	0.608956	1681.629	0.735964
10.0.24.1	559361	1258	0.002249	382380416	512000	746.83675	2018-May-27	0.303186	1723.80403	0.723044
10.0.15.1	107305	71	0.000662	86219616	512000	168.397688	2018-May-27	0.609522	1650.4475	0.735724
10.0.7.1	107228	134	0.00125	85982064	512000	167.933719	2018-May-27	0.506492	1650.681	0.731587
10.0.11.1	220294	264	0.001198	165032000	512000	322.328125	2018-May-27	0.507502	1671.95775	0.732613
10.0.23.1	220035	643	0.002922	165511808	512000	323.26525	2018-May-27	0.305653	1651.06475	0.722703
10.0.9.1	110105	105	0.000954	82247872	512000	160.640375	2018-May-27	0.55567	1703.978	0.733539
10.0.29.1	109998	104	0.000945	80708096	512000	157.633	2018-May-27	0.507502	1704.16	0.731443
10.0.36.1	219606	505	0.0023	159117504	512000	310.776375	2018-May-27	0.302724	1711.95775	0.722371
10.0.37.1	113402	405	0.003571	83273344	512000	162.64325	2018-May-27	0.305653	1806.388	0.721792
10.0.21.1	111480	419	0.003759	78643040	512000	153.599688	2018-May-27	0.201511	1798.1795	0.715065

Figura 42- Exemplo de um *dataset* usado no cálculo do índice.

5.3 Análise das fontes de dados utilizados

Tal como aconteceu anteriormente, tivemos que verificar se o *dataset* agora preparado têm os vários elementos de dados com um nível de qualidade aceitável para suportarem o processo de

cálculo do índice. Na prática, basicamente, verificámos se todos os registos podiam ser usados no processo e se as suas falhas, caso existam, podem ser recuperadas. Além disso, depois preparados e limpados, os dados contidos no *dataset* têm de ser mapeados de forma a poderem povoar uma nova estrutura multidimensional de dados. Foi a partir da análise deste *dataset* que pudemos identificar as várias dimensões que iremos utilizar no *data mart* que concebemos para o acolhimento dos valores do índice de monitorização calculados ao longo do tempo. Mais uma vez foi necessário proceder ao desenvolvimento do mapeamento lógico dos dados (Tabela 11). Depois de reunir os fatores que consideramos importantes para o cálculo do índice, desenvolver um mapeamento lógico, ficou relativamente mais simples. Uma vez que os dados presentes no *dataset* foram sintetizados em laboratório, não ficou muita margem para erro, por isso os dados presentes no *dataset* não precisam de sobre nenhuma alteração para serem carregado para o novo *data mart.*

Tabela 11- Mapeamento lógico dos dados usado para o cálculo do índice.

	Target				source		
database	table	table type	column	database	column	datatype	Transfomation
dw_Index	DimLigacao	dimension	nome	diaIndex*.csv	NomeLigação	String	direto
dw_Index	DimLigacao	dimension	idLigacao				Surrogate Key
dw_Index	ftIndex	factTable	tamanhoTotal	diaIndex*.csv	TamanhoTotal	Float	direto
dw_Index	ftIndex	factTable	totalPacotes	diaIndex*.csv	TotalPacotes	Inteiro	direto
dw_Index	ftIndex	factTable	totalPacotesLost	diaIndex*.csv	TotalPerdidos	Inteiro	direto
dw_Index	ftIndex	factTable	velocidadeLinha	diaIndex*.csv	Velocidade	Float	direto
dw_Index	ftIndex	factTable	tempo	diaIndex*.csv	TempoNomes	Float	direto
dw_Index	ftIndex	factTable	percentagemPerdidos	diaIndex*.csv	PercentagemPerdidos	Float	direto
dw_Index	ftIndex	factTable	tempoTransporte	diaIndex*.csv	TempoTransporte	Flaot	direto
dw_Index	ftIndex	factTable	tempoProcessamento	diaIndex*.csv	TempoProcessamento	Float	direto
dw_Index	ftIndex	factTable	valorIndice	diaIndex*.csv	ValorIndice	Float	direto
*O nome do ficheiro csv é referente aos vários dias. Exemplos Segunda, Terça, etc.							

Como é possível observar na Figura 42, apenas duas colunas são possíveis de serem organizadas de modo a que seja possível fazer operações sobre esses dados. Essas colunas são: a coluna tempo e a coluna *NomeLigacao*. A coluna tempo não interessa para o mapeamento visto que será carregada por um ficheiro CSV a parte. Já a coluna *NomeLigacao* é onde estão guardados todos os endereços de IP importantes para a qualidade da rede, logo é importante que estes registos estejam guardados de forma organizada e identificados apenas por um único identificador. Logo é fácil de perceber o motivo dos endereços de IP serem caracterizados como dimensões no *data mart* idealizado.

Com a sintetização dos dados utilizados o nosso único objetivo era conseguir responder a estas duas perguntas:

- Qual o índice de qualidade de uma dada ligação num determinado dia?
- Qual o índice final da rede?

Estas perguntas permite-nos um acompanhamento sobre a qualidade da rede nos seus vários pontos. Desta maneira ficamos a saber quais são os pontos da rede que apresentam maior debilidades, podendo intervir de maneira mais rápida e assertiva possível.

5.4 A Estrutura de Dados de Suporte ao *Data Mart Index*

Uma vez analisado o sistema de dados chegou a altura de desenvolver um sistema multidimensional que permita acolher os dados registados. Novamente recorremos ao método dos "4 passos" para desenvolvemos o sistema multidimensional. A estrutura desenvolvida não apresenta um grau de complexidade elevado visto que é apenas um (pequeno) *data mart*, onde será guardado toda a informação relativa à qualidade do sistema.

Tabela 12-Matriz de decisão do *Data Mart Index*

	Caraterização de *Data Mart Index*	
	Identificação: Cálculo do Índice	
	Descrição Geral: Informação de suporte para o cálculo do índice em cada ponto da rede num determinado dia	
	Estrutura Base	
	Tabela de Factos	*ftIndex*
	Dimensões	
	DimData	**X**
	DimLigacao	**X**
	Número de Dimensões	2
	Tipo	Transacional
	Periodicidade	Diária
	Descrição	Índice em cada ponto da rede.
	Utilidade Estratégica	Melhorar a rede disponível.

Utilizadores	Gestores de Rede
Observações	Nada a assinalar

Assim, começámos por definir a nossa matriz de decisão relativa ao *data mart* - "*Index*" – que queremos implementar (Tabela 12). A matriz de decisão é bastante simples visto apresentar apenas duas dimensões.

Uma vez terminada a matriz de decisão chegou a altura de descrever as várias dimensões presentes no *data mart* "*index*". As dimensões presentes no *data mart* "*index*" são as seguintes:

- Dimensão Data (*DimData)*, data ao qual foi retirada a informação relativa as várias ligações presentes na rede. Na Tabela 13 é possível observar a caracterização da dimensão data para o *data mart index.*

Tabela 13-Caracterização da dimensão data no *data mart index.*

Caracterização da dimensão					
Identificação			*DimData*		
Descrição			Calendário do ano e os seus atributos.		
Tipo			Sem variação		
Crescimento			Não cresce. O povoamento desta dimensão é feito durante a fase de arranque do *Data Warehouse* para um período de 2 anos, desde a data mais antiga, i.e., 2018-01-01.		
Atributos					
Nr	**Identificação**	**Variação [Sim/Não]**	**Domínio**	**Descrição**	**Exemplo**
1	idData	N	Inteiro	Número único que identifica uma determinada data	1
2	Data	N	Data	Data do calendário	2018-08-30
3	Dia	N	Inteiro	Número do dia do Mês	30
3	Mês	N	Inteiro	Número do Mês	8
4	Semana	N	String	Nome do dia da Semana do ano	Segunda
5	Trimestre	N	Inteiro	Semestre em que o mês se refere	3
7	Ano	N	Inteiro	Ano da data	2018

Hierarquia(Ramos)		
Nr	**Identificação**	**Esquema**
1	H1	idData->data->mês->trimestre->ano->All
2	H2	idaData->data->dia->All
3	H3	idData->data->diaSemana->All
Perfis de Utilização		
Gestores de rede e administradores.		
Observações		
Nada a acrescentar		

- Dimensão Ligação (*DimLigacao*), representa todas as ligações IP presentes na rede. Na Tabela 14 é possível observar uma caracterização geral da dimensão ligação.

Tabela 14-Caraceterização da dimensão ligação para o *data mart index.*

Caracterização da dimensão					
Identificação			*DimLigacao*		
Descrição			Representa todos os endereços de IP presentes na rede.		
Tipo			Sem Variação		
Crescimento			0.10% dia.		
Atributos					
Nr	**Identificação**	**Variação [Sim/Não]**	**Domínio**	**Descrição**	**Exemplo**
1	idLigacao	N	Inteiro	Representa o identificador de uma ligação.	1
2	Nome	N	String	Representa o nome da ligação	10.0.0.20
Hierarquia(Ramos)					
Nr	**Identificação**	**Esquema**			
1	H1	idLigacao->Ligacao->All			
Perfis de Utilização					
Gestores de rede e administradores.					
Observações					
Nada a acrescentar					

Posto a descrição das várias dimensões, assim como a apresentação da matriz de decisão, escolher o grão a incorporar no *data mart index,* foi bastante simples. O grão escolhido vai de encontro as nossas necessidades, logo só poderia ser o valor do índice para num determinado dia para uma

determinada ligação. Posto isto a Tabela 15 apresenta um resumo das várias dimensões usadas para o *data mart index.*

Tabela 15-Síntese das Dimensões

Dimensões *Data Mart Index*			
Nr	Identificação	Descrição	Esquema(Tipo)
1	Ligação	Nome da ligação	DimLigacao (Sem variação)
2	Data	Dimensão temporal. Acolhe todos os atributos ao longo do tempo, como data, dia da semana, mês, trimestre e ano	DimData (Sem variação)

Como forma de resumir o processo de modelação dimensional utilizámos, novamente, o esquema dimensional produzido para o *Data Mart Index* (Figura 43), realizado com base na notação de *Golfarelli et al.* [5], usando a ferramenta *draw.io* [18].

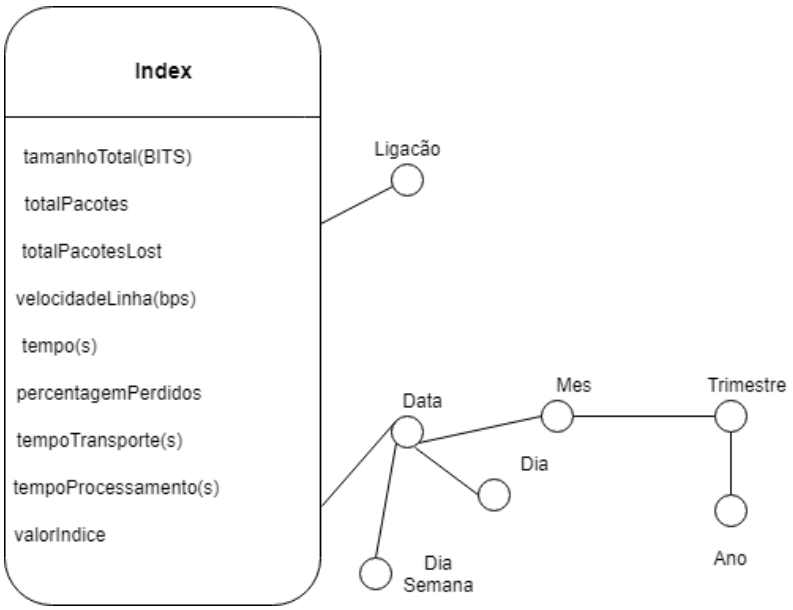


Figura 43-Esquema estrela para o acolhimento dos registos do índice.

A tabela de factos *"Ft_Index"* (Tabela 16) escolhida permite visualizar a qualidade do serviço diariamente. Para tal ela acolhe uma série de registo que são importantes no cálculo do índice de bem-estar do serviço. Os registos também permitem ter uma noção mais alargada da rede e que alterações devem ser feitas para melhorar a qualidade do serviço.

Tabela 16- Caracterização da tabela de factos *"Ft_Index".*

Caracterização da tabela de factos	
Identificação	Ft_Index
Descrição	Tabela que acolhe o valor do índice.
Data *mart*	Comercial
Tipo	Transacional
Utilidade estratégica	Acompanhar a qualidade de cada ligação ao longo dos vários dias
Povoamento	Realizado diariamente entre as nove horas e onze horas da noite.
Dimensão inicial	
Crescimento	0.10% dia.
Período de dados	Desde o ano de 2018. Os anos anteriores ficarão em arquivos.
Atributos	
Dimensões	

Nr	**Identificação**	**Chave**	**Domínio**	**Descrição**	**Exemplo**
1	IdLigacao	S	Inteiro	Código interno do nome da ligação	1
2	IdData	S	Inteiro	Código da data referente a data em que o pacote circulou na rede.	1

Medidas				
Nr	**Identificação**	**Domínio**	**Descrição**	**Exemplos**
1	tamanhoTotal	Float	Somatório dos vários pacotes que circulam numa ligação	2
2	totalPacotes	Inteiro	Total de pacotes que circulam numa ligação	10
4	totalPacotesLost	Inteiro	Total de pacotes perdidos	10

			numa ligação	
5	velocidadeLinha	Float	Velocidade de uma dada ligação	10
6	tempo	Float	Tempo que uma dada ligação demora a estabelecer uma conexão	0.4
7	percentagemPerdidos	Float	Razão entre os pacotes perdidos e os pacotes totais	0.1
8	tempoTransporte	Float	Tempo que uma dada ligação demora a despachar os vários pacotes que nela circulam	10
9	tempoProcessamento	Float	Tempo médio que uma dada ligação demora a processar um determinado pacote	1
10	valorIndice	Float	Valor do índice de bem-estar de uma dada ligação	0.1

Índice

Nr	Identificação	Tipo	Descrição
1	IdIndice	Primário	Único, ordenado fisicamente(*clustered*) de forma crescente.
2	IdData	Secundário	Ordenado de forma crescente.
3	idLigacao	Secundário	Ordenado de forma crescente.

Perfis de Utilização

Administrador da base de dados e gestores de rede

Observações

Todos os valores temporais considerados nos atributos temporais estão em segundos, se estes valores estiverem em outras unidades temporais é necessário fazer a conversão primeiro. O atributo "*tamanhoTotal*" está em bits, qualquer unidade de informação que não seja o bit será necessária proceder a sua conversão. A velocidade da linha está em bits por segundo(*bps*) qualquer unidade de transmissão de dados que não seja o *bps* é necessário proceder a sua conversão.

A construção do *data mart index* têm como objetivo ajudar-nos a tomar melhores decisões para uma melhoria do tráfego da rede. Para percebemos melhor como se relacionam as tabelas de dimensões com a tabela de factos, recolhemos a Figura 44, onde é apresentado um esquema lógico sobre o *data mart index.*

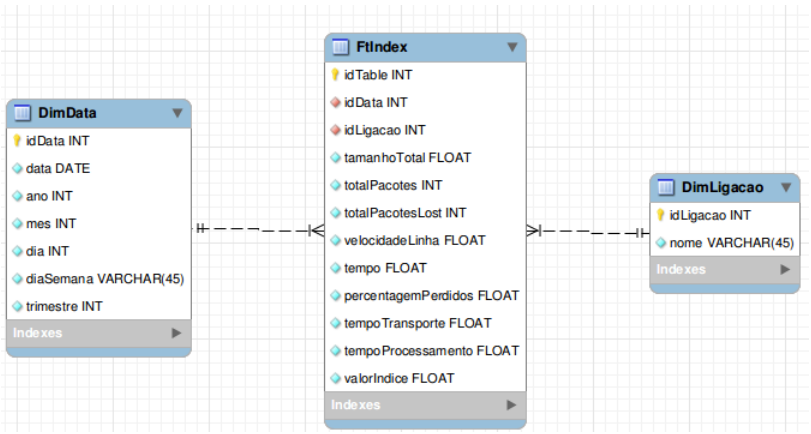


Figura 44-- Esquema logico para o *Data Mart "Index".*

O esquema lógico apresentado é baste simples, visto possuir apenas uma tabela de factos e duas dimensões. No entanto ele acolhe as informações necessárias para o cálculo do índice de qualidade da rede, nos vários pontos de ligação.

Uma vez traduzido o esquema lógico para um esquema físico, permite-nos responder algumas questões sobre a qualidade da rede utilizada (Figura 45).

```
select nome,dia, valorIndice
from dimData,ftIndex,dimLigacao
where dimData.idData=ftIndex.idData
and dimLigacao.idLigacao=ftIndex.idLigacao
```

Figura 45-Valor do índice por dia em cada ligação.

Esta pergunta retorna por dia o valor do índice de qualidade em cada ligação da rede, conseguindo desta maneira identificar mais rapidamente os pontos problemáticos da rede.

5.5 O Povoamento da Estrutura Multidimensional do *Data Mart Index*

Após a definição do modelo dimensional do *data warehouse* é necessário proceder ao seu povoamento. Começamos então por identificar os vários tipos de carregamentos associados. Mais uma vez irão ser dois tipos de carregamentos, o carregamento inicial e o regular. O carregamento inicial que apenas é efetuado uma vez e se destina ao povoamento inicial das tabelas de dimensões e da tabela de factos. O carregamento regular irá ser o responsável por adicionar novos elementos ao *data warehouse.*

O povoamento do *data mart "índex"* não é tão complicado como o do *data mart "trafego"*. Começamos por extrair a informação presente nos ficheiros *CSV* (Figura 46). Mais uma vez a dimensão Data foi carregada por um ficheiro *CSV* gerado a parte.

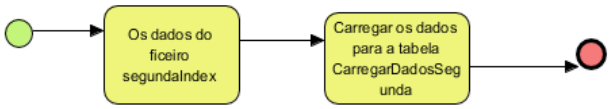


Figura 46-Extração dos ficheiros para o cálculo do índice.

Posteriormente é necessário gerar novas chaves para as várias ligações presentes na rede (Figura 48). Uma vez esse processo estar concluído os vários dados serão inseridos nas tabelas correspondentes (Figura 47).

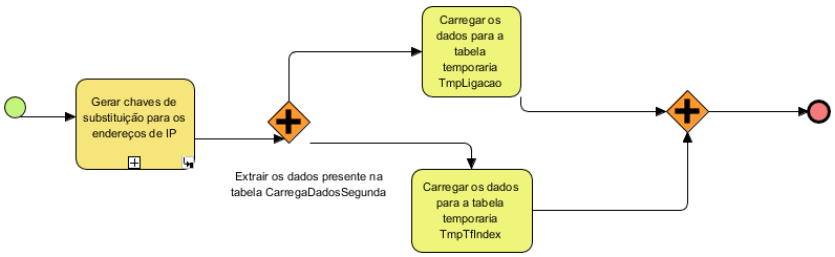


Figura 47-Inserção dos registos nas tabelas temporárias do *ETL index.*

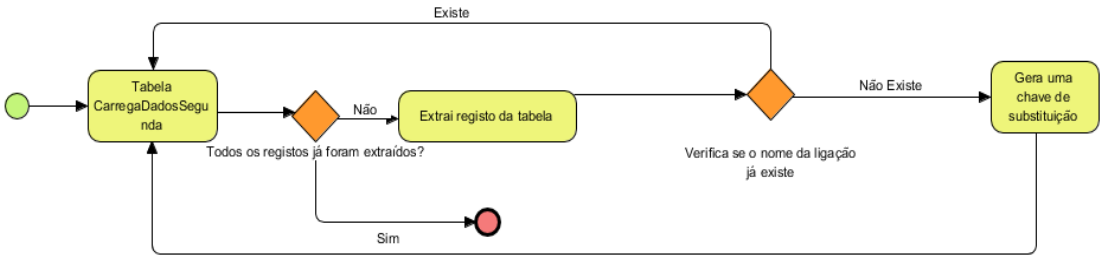


Figura 48- Subprocesso de geração de chaves para os endereços de IP do *data mart index*.

Feito o processo de carregamento para as tabelas temporárias fica a faltar o carregamento para o *data mart index*. Mais uma vez é preciso, primeiramente, carregar os dados presente na tabela de dimensão temporária(*TmpDimLigacao*) para a dimensão correspondente(*DimLigacao*) para depois procedemos ao carregamento da tabela de factos (Figura 49).

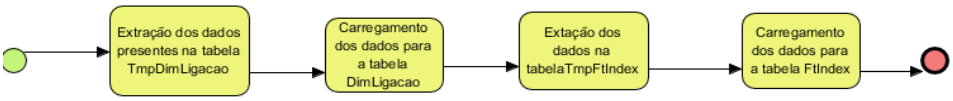


Figura 49- Processo de carregamento para o *data mart index.*

5.6 Implementação do Processo de Povoamento para o *Data Mart Index*

Feito uma descrição geral sobre o processo de povoamento do *data mart index* chegou a altura de falar sobre o seu processo de implementação. Mais uma vez a ferramenta usada foi o *Pentaho Data Integration* (*Kettle*) [14]. Como já referido existem dois tipos de povoamento, o povoamento inicial, que é feito no arranque do sistema, e o povoamento regular, que é feito com o passar do tempo, sendo feito um refrescamento à posteriora da cache da base de dados multidimensional. No povoamento inicial (Figura 50), as datas, geradas num ficheiro CSV, são carregadas para a tabela dimensão (*DimData*) presente no *data mart index* e também para as tabelas de *surrogate key* presentes no ETL. Quando este processo estiver concluído, carregamos os dados referentes ao

cálculo do índice de qualidade para uma base de dados *mysql* a fim de eles serem analisados, tratados e depois inseridos no *data mart index*.

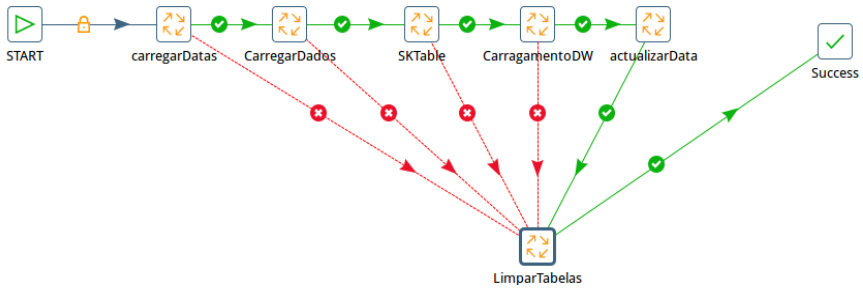


Figura 50-Vista geral sobre o povoamento inicial do *data mart index.*

O carregamento regular (Figura 51) é muito semelhante ao carregamento inicial, só que desta vez já não é carregado as datas presentes no ficheiro CSV e no fim de os dados serem inseridos no *data mart index* é necessário proceder a uma atualização da cache da base de dados multidimensional, para conseguirmos ter acesso aos novos registos.

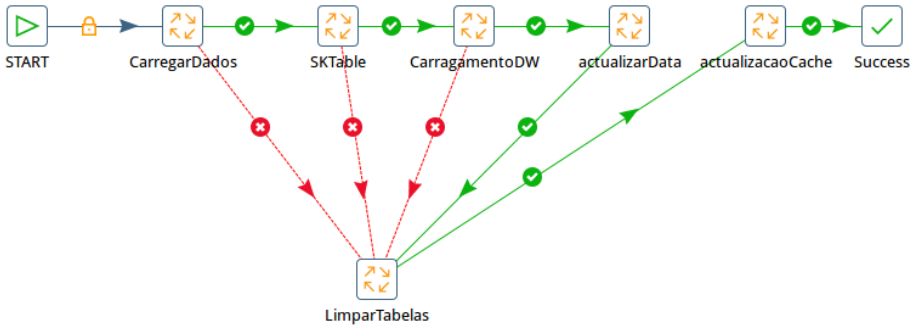


Figura 51- Vista geral sobre o povoamento regular do *data mart index.*

Para ajudar na inserção dos elementos recolhidos nas tabelas temporárias foi necessário recorrer, novamente, a um *stored procedure* (Figura 52). O *stored procedure* é muito útil para evitar problemas de desnormalizarão. O *stored procedure* gera chaves de substituição para os vários endereços presentes na rede e faz a associação dos instantes temporais presentes nos ficheiros

CSV, sintetizados para o cálculo do índice de qualidade, com as datas presentes no sistema. Mais tarde o *stored procedure* carrega os registos para a tabela temporária de factos.

```
start transaction;
select count(*) into quantidade
from SkRouter
where nome=in_nomeRouter;
if quantidade=0
then
insert into SkRouter (idRouter,nome,dataInsert) value(idRouter,in_nomeRouter,now());
end if;
select idRouter into id
from SkRouter
where nome=in_nomeRouter;
select idData into idDataTmp
from SkData
where SkData.data=in_date;
select count(*) into quantidade
from DimRouter
where routter=in_nomeRouter;
if quantidade=0
then
insert into DimRouter (id,routter) value(id,in_nomeRouter);
end if;
```

Figura 52- Amostra do *stored procedure* usado para a geração de chaves de substituição.

5.7 A Base de Dados Multidimensional *Index*

Despois de definida a estrutura do *data mart index* passámos à criação correspondente da base de dados multidimensional. Para tal, mais uma vez, utilizámos a ferramenta *Pentaho* [10].

```
<Cube name="indexCube" visible="true" cache="true" enabled="true">
  <Table name="facttable">
  </Table>
  <DimensionUsage source="DimLigacao" name="DimLigacao" visible="true" foreignKey="idLigacao" highCardinality="false">
  </DimensionUsage>
  <DimensionUsage source="DimData" name="DimData" visible="true" foreignKey="idData" highCardinality="false">
  </DimensionUsage>
  <Measure name="tamanhoTotal" column="tamanhoTotal" datatype="Numeric" formatString="#" aggregator="sum" visible="true">
  </Measure>
  <Measure name="totalPacotes" column="totalPacotes" datatype="Integer" formatString="#" aggregator="sum" visible="true">
  </Measure>
  <Measure name="totalPacotesLost" column="totalPacotesLost" datatype="Integer" formatString="#" aggregator="sum" visible="true">
  </Measure>
  <Measure name="velocidadeLinha" column="velocidadeLinha" datatype="Numeric" formatString="#" aggregator="distinct-count" visible="true">
  </Measure>
  <Measure name="valorIndice" column="valorIndice" datatype="Numeric" formatString="#.######" aggregator="distinct-count" visible="true">
  </Measure>
  <Measure name="tempo" column="tempo" datatype="Numeric" formatString="#.######" aggregator="distinct count" visible="true">
  </Measure>
</Cube>
```

Figura 53- Fragmento de um ficheiro XML com a definição de estrutura de dados multidimensional *index*.

Para procedemos à criação da base de dados multidimensional *index* foi necessário, novamente, converter a base de dados para num formato *XML* (Figura 53) e depois fazer a sua importação para *Pentaho* [10]. Fazer a conversão de uma base de dados para o formato *XML* sem ajuda de nenhuma ferramenta é possível e requer algum trabalho, no entanto visto que existe uma ferramenta própria que trata desse tipo de conversões é mais aconselhável o uso desta ferramenta, para diminuir o número de erros ocorridos durante a conversão. A ferramenta utilizada é, novamente, o *Pentaho Schema Workbench* [13] e podemos encontrar uma imagem do seu ambiente com a base de dados multidimensional *índex* já contruída na Figura 54.

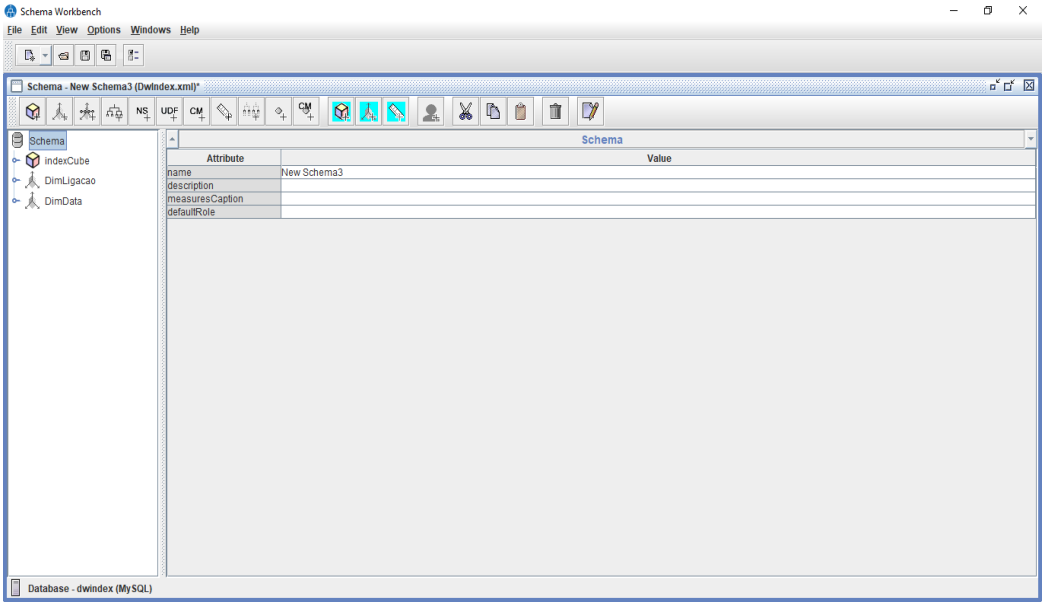


Figura 54- O ambiente da ferramenta Schema Workbench da estrutura multidimensional *index.*

Depois de analisar a estrutura do *data mart index* a estrutura do cubo fica mais fácil. O cubo irá conter duas dimensões que são *"DimData"* e *"DimLigacao"*. Todas estás dimensões são retiradas diretamente do *data mart index* já implementado. Em relação as medidas irão ser as mesmas apresentados no *data mart index* que são as seguintes: *"tamanhoTotal", "totalPacotes","totalPacotesLost","velocidadeLinha","tempo","percentagemPerdidos", "tempoTransporte", "tempoProcessamento"* e o *"valorIndice"*.

No que diz respeito as hierarquias que compõem cada dimensão tem-se:

DimData

H1: idData->data->mês->trimestre->ano->All

H2: idaData->data->dia->All

H3: idData->data->diaSemana->All

DimLigacao

H1: idLigacao->Ligacao->All

Capítulo 6

Análise do Sistema de Monitorização

6.1 Visualização de dados

Os *dashboards* [19,33] são uma das várias ferramentas de visualização de dados que permitem exibir o estado atual de um conjunto de métricas e de indicadores chaves de desempenho (*KPIs*) para um dado sistema de análise. Devido às diversas áreas em que podem ser adaptados de modo a exibir métricas específicas orientadas a um determinado ponto de vista ou área de negócio em particular [19]. Qualquer que seja a utilização de um determinado *dashboard*, este deve possuir um conjunto de características fundamentais, entre as quais se salientam [19,33]:

- **Simples**, permitindo comunicar com clareza os dados para o qual foi construído.
- **Focado**, apresentando informação bem orientada, não se desviando dos seus princípios e não distraindo o utilizador com outra informação.
- **Organizado em termos da sua informação** de negócio de forma a preservar o seu significado e a sua utilização.
- **Atual** em relação aos últimos meios de compreensão em termos de perceção humana relativos à apresentação visual da informação.
- **Agradável à vista,** disponibilizando meios de visualização de dados atrativos e com design adequado aos seus utilizadores**.**

Relativamente aos tipos de *dashboards* existentes, estes podem ser divididos em três grandes grupos [19]:

- ***Dashboards* estratégicos**, que monitorizam o progresso de um dado sistema ou processo para atingir metas pré-definidas, e acompanham a visualização dos dados com *KPI's* relevantes.
- ***Dashboards* analíticos ou táticos** que, ao contrário dos *dashboards* estratégicos, são preparados para propósitos mais detalhados, sendo usados geralmente para traçar as tendências em relação aos objetivos e iniciativas de um dado sistema.
- ***Dashboards* operacionais** que monitorizam atividades ou variáveis que necessitam de um acompanhamento em tempo real, fornecendo informações detalhadas.

A distinção entre cada tipo de *dashboards* não resulta apenas das diferentes funções que cada um apresenta, mas também da audiência a que cada um se destina.

Existem, também, algumas "regras" que devem ser tidas em atenção antes da conceção de um determinado *dashboard*, a destacar [19]:

1. Identificar quem é que se pretende impressionar;
2. Selecionar o tipo mais correto de *dashboard*.
3. Agrupar os dados de forma lógica, fazendo uma utilização sensata do espaço;
4. Fazer com que os dados sejam relevantes para a sua audiência.
5. Organizar da melhor forma o *dashboard* apresentando apenas as métricas mais importantes.
6. Determinar a frequência com que os dados precisam de ser realmente refrescados.

O correto cumprimento destas regras fará com que qualquer *dashboard* se torne num produto útil para as partes interessadas, capaz de responder às suas necessidades.

6.2 Implementação final das *dashboards*

Os *dashboards* que foram idealizados e implementados são alimentados a partir de um conjunto de vários cubos, tendo sido concebidos com o objetivo concreto de auxiliar os seus utilizadores na análise de qualidade de serviço de uma rede. Com a informação disponibilizada pelos *dashboards* é possível retirar algumas indicações bastante concretas sobre o que se está a passar em determinado momento da rede e analisar com detalhe as causas que estão por trás deste ou

daquele evento. Em concreto, neste trabalho de dissertação desenvolveram-se dois *dashboards* distintos, nomeadamente:

1. Para análise do tráfego do sistema de rede.
2. Para análise da informação do índice de monitorização estabelecido e a sua evolução.

O primeiro destes *dashboards* for implementado especificamente para suportar a visualização da informação relativa ao tráfego da rede e informação subjacente. Este *dashboard* incorpora os seguintes elementos (Figura 56):

1. Um gráfico de barras, para visualização do número de pacotes perdidos por dia na rede.
2. Um gráfico de barras, para revelar o tráfego da rede em cada dia de serviço.
3. Um gráfico de barras, para permitir ao utilizador perceber quais os dias da semana nos quais existem mais tráfego na rede.
4. Um gráfico de barras, para possibilitar ao utilizador perceber quais são as horas nas quais mais tráfego circula.
5. Um gráfico de barras, para ajudar o utilizador a perceber quais os pontos de ligação por onde mais tráfego circula.

Este *dashboard* é alimentado a partir da estrutura multidimensional "*TrafegoCube*", que é referente ao *data warehouse* que armazena o tráfego resultante da rede. A alimentação do *dashboard* é feita através dos dados presentes no cubo "*TrafegoCube*", que por sua vez é alimentado através do *data warehouse* referente. Quando se dá a inserção de novos dados no *data warehouse*, o cubo não tem imediatamente acesso aos mesmos, sendo necessário atualizar a cache do sistema. Aproveitando o facto de o *Pentaho Business Analytics* [10] disponibilizar um serviço web para limpeza de cache, procedemos à sua invocação (Figura 55). Sempre que houver uma inserção de novos dados no ETL, este serviço é invocado e faz com que a cache do cubo se atualize.

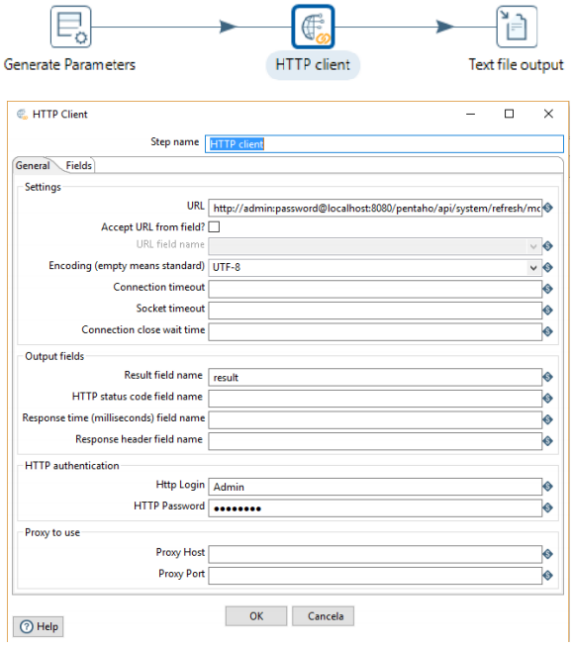


Figura 55-Invocação HTTP para limpeza da cache.

Figura 56-*Dashboard* para análise do tráfego da rede.

Assim, pelo *dashboard* criado (Figura 56), podemos verificar o acesso a um conjunto de informações que lhe nos permite avaliar o estado atual da rede. Através da informação apresentada conseguimos retirar conclusões sobre o estado da rede. Sabendo o número de pacotes que foram perdidos na rede e o tráfego gerado por dia, permite-nos perceber se com o aumento do tráfego se traduz num maior número de pacotes perdidos. Através da informação do

tráfego gerado por semana, conseguimos avaliar quais os dias da semana onde existem mais utilizadores na rede, podendo reforçar a rede para esses dias da semana. Através do tráfego gerado por hora conseguimos saber quais são os períodos do dia que mais utilizadores utilizam a rede e assim minimizar as quebras da rede nesses períodos de tempo. Sabendo quais as ligações por onde circulam mais tráfego na rede ficamos a perceção sobre quais os pontos que devem ser reforçados para minimizar as quebras na rede.

Quanto ao segundo *dashboard* que foi implementado, aquele que permite analisar a cada momento o índice da qualidade de serviço da rede (Figura 57), este é constituído por um conjunto de elementos um pouco diferentes do *dashboard* anterior. A sua estrutura incorpora dois elementos visuais, nomeadamente:

1. Um gráfico de barras, que permite ao utilizador visualizar o índice de bem-estar da rede em cada dia.
2. Vários gráficos de linhas que irão permitir ao utilizador visualizar o índice de bem-estar de cada ligação nos vários dias da semana.

Este *dashboard* é alimentado a partir da estrutura multidimensional "*IndexCube*", que é referente ao *data warehouse* que armazena o índice de bem-estar da rede. A alimentação do *dashboard* é feita através dos dados presentes no cubo "*IndexCube*", que por sua vez é alimentado através do *data warehouse* referente. O processo de refrescamento do *dashboard* (Figura 55) é igual ao processo de refrescamento do *dashboard* anterior (Figura 56).

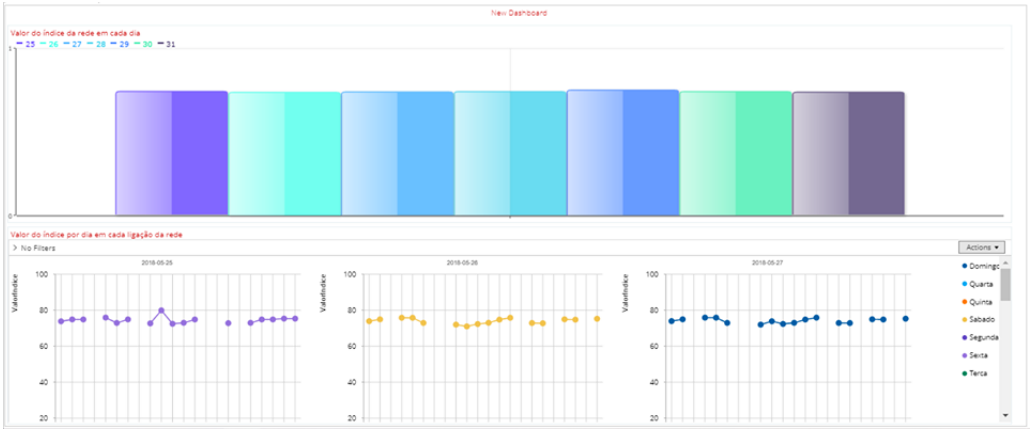


Figura 57-*Dashboard* para análise do índice de monitorização.

O *dashboard* do índice de monitorização (Figura 57) verifica os diferentes valores médios do índice de bem-estar ao longo dos vários dias. Com isso é possível saber a qualidade da rede para cada dia e com a ajuda do *dashboard* anterior (Figura 56) conseguimos ter uma perceção sobre os motivos desse índice. O *dashboard* do índice de monitorização também permite perceber quais os pontos da rede que pior desempenho trazem à rede, assim é possível perceber quais os pontos que devem ser melhorados.

De facto, a utilização de sistemas de análise de dados suportados por *dashboards* é algo que ajuda a compreender de uma forma simples, mas bastante efetiva, o estado ou comportamento de um dado sistema ou processo, num dado período de tempo, de acordo com um conjunto de métricas a parâmetros bem-estabelecidos. No caso de estudo que trabalhámos ao longo desta dissertação, isso pode ser comprovado pela facilidade com que fizemos a análise e a sustentação do serviço prestado pela rede que idealizámos. Porém, essa facilidade só é possível se estabelecermos as estruturas e os mecanismos de análise à priori, com uma forte fundamentação e suporte de dados, implementados num sistema apropriado. Isso foi conseguido neste trabalho graças a um planeamento prévio sobre o armazenamento e visualização da informação. A basta informação acolhida no *data warehouse* permite uma maior certeza sobre a qualidade da rede. A identificação dos *outliers* presentes na rede de *process mining* foram descartados, dando assim uma informação mais fidedigna sobre a qualidade da rede.

Capítulo 7

Conclusões e Trabalho Futuro

7.1 Comentários Finais

Apesar de ser ainda pouco utilizado nos dias de hoje, o *process mining* revela grandes potencialidade na análise de processos de negócio e dos eventos que eles se relacionam, algo que usualmente tem um grande impacto na forma como as empresas estão organizadas e como é que poderão melhorar o seu desempenho. A aplicação de técnicas de *process mining* não só permite suportar melhores processo de análise de dados, como também verificar a forma como determinados processos estão a decorrer através da análise da sua evolução ao longo do tempo. É possível aplicar técnicas de *process mining* em várias áreas aplicacionais, como é o caso da banca, da saúde, ou até mesmo em processos de otimização de desempenho de uma empresa.

Neste trabalho de dissertação aplicamos um conjunto de técnicas de *process mining* para analisar o comportamento de uma rede de comunicações por comutado idealizada por nós e criada usando a ferramenta *CORE Network Emulator.* Através do sistema simulado e dos dados recolhidos ao longo de um período de tempo de trabalho da rede foi possível verificar o comportamento da referida rede na satisfação dos vários pedidos de serviço verificados, sabendo-se quais os caminhos que os pacotes percorriam até chegar ao seu destino. Data a rede ser um "objeto" de laboratório, não nos foi possível aplicar o trabalho realizado na análise do tráfego de uma aplicação

real. Assim, tivemos de simular o tráfego eventualmente ocorrido entre os vários dispositivos da rede e captar a informação relativa aos dispositivos envolvidos nos processos de comunicação. Por vezes, o processador do computador que utilizámos como base de trabalho não aguentou o esforço requerido pelo processo de simulação, acabando por desligar. Isso dificultou um pouco o trabalho de preparação dos dados e, consequentemente, da sua posterior análise. Depois da análise estar concluída foi necessário combinar os vários ficheiros obtidos através do processo de recolha para um único recipiente (um ficheiro específico) de forma a que este fosse possível ser analisado pela ferramenta de mineração- o *Disco.* Devido a termos tido acesso a uma licença para estudantes, esta ferramenta não permitiu realizar processos de análise que envolvessem um volume de dados superior a cinco milhões de registos. Como tal, foi necessário proceder a separação física do ficheiro inicialmente obtido em vários outros ficheiros (*chunks*), de forma a podermos processar todos os dados que conseguimos sintetizar. Uma vez concluído o processo de processamento dos vários ficheiros chegamos à altura de exportar os dados da análise fornecidos pela ferramenta. Uma vez retirados esses ficheiros de análise procedemos à criação do ficheiro que iria ser fundamental no cálculo do índice de qualidade. Uma vez o ficheiro ter sido desenvolvido chegou a altura de carregar os ficheiros para os dois *data marts* desenvolvidos. Primeiramente carregamos toda a informação relativa ao tráfego gerado por uma rede, visto que estes ficheiros continham milhões de registos, a base de dados não consegui carregar os registos todos, então procedemos a partição desses ficheiros e carregamos os registos para a base de dados. Como a base de dados multidimensional utiliza uma forma diferente de representação dos números decimais, no início do processo, todos os números eram arredondados às unidades. Foi necessário modificar a representação destes valores para que a base de dados multidimensional deixasse de arredondar os valores as unidades.

O desenvolvimento desta dissertação permitiu consolidar os nossos conhecimentos acerca da importância que uma boa rede informática apresenta nos dias de hoje. Ao longo do processo de desenvolvimento da dissertação ficamos a conhecer um pouco melhor a área de rede de comunicação de redes. Consolidamos a nossa perceção sobre o real funcionamento de uma rede de comunicação e os vários fatores que provocam o atraso nas comunicações. Ao aplicar técnicas de *process mining* reforçamos os nossos conhecimentos sobre exploração, tratamento e processamento da informação recolhida. O conhecimento sobre base de dados multidimensionais também sai reforçado graças a criação dos vários modelos multidimensionais.

Como nota final, podemos dizer que com a realização deste trabalho foi possível ter uma perceção bastante concreta do impacto que uma rede de comunicações por computador pode ter em qualquer local de trabalho. A realização regular de processos de monitorização dos serviços de uma rede por parte das instituições pode evitar quebras de serviço, bem como prevenir potenciais situações relacionadas com ataques informáticos. De facto, é importante que as instituições tenham um bom conhecimento da qualidade de serviço das suas redes para conseguirem atuar quando estas não funcionem como seria expectável. A sua existência de um sistema para a monitorização da qualidade de serviço de uma rede, baseado em índices de qualidade, como aquele que foi desenvolvido e validado neste trabalho, é de facto um elemento de bastante valor acrescentado para qualquer instituição.

7.2 Próximos Passos

Neste trabalho de dissertação demonstrámos a forma como poderíamos utilizar *process mining* para analisar a qualidade de serviço de uma rede e sustentar a criação e manutenção de um índice de "bem-estar" para a própria rede, algo que nos permitisse ver, de uma forma simples, mas clara, o estado corrente da rede num determinado período. Todavia, os dados que utilizámos para sustentar a abordagem, modelos e sistemas implementados tiveram que ser sintetizados por nós em laboratório, através da criação de um programa que representasse e simulasse o comportamento de uma dada rede desenhada por nós. Como tal, apesar de termos demonstrado a utilização das técnicas de *process mining* neste tipo de aplicações não foi possível, por motivos vários, validar o sistema implementado numa situação do mundo real.

Assim, um dos próximos passos passaria por testar esta solução agora encontrada num ambiente real, coletando e tratando, se possível em tempo-real, a informação relativa ao desempenho do sistema de rede e revelar o seu índice de qualidade de serviço, a cada momento, ao longo do tempo de funcionamento da rede. Outro aspeto interessante, seria explorar novas vertentes de análise, novos elementos de dados, para melhorar (afinar) a fórmula do cálculo do índice e, consequentemente, torna-lo mais rico e mais ajustado a cada cenário de aplicação em análise. De facto, seria bastante curioso analisar a segurança dos dispositivos incorporados numa rede, inspecionando, por exemplo, se um router, é ou não vulnerável a ataques informáticos. Além

disso, com a implementação do sistema de análise numa aplicação real poder-se-ia fazer um ajustamento mais adequado aos pesos atribuídos a cada um dos elementos que entra no cálculo do índice de qualidade de serviço da rede.

Bibliografia

[1] Combs, Gerald "Wireshark" [Online]. Available: https://www.wireshark.org/. [Accessed:04-09-2018].

[2] (van der Aalst 2011)
Aalst, Wil M. P. van der. 2011. "Process Mining." *Process Mining*.

[3] P.M.G. Eindhoven Technical University, "ProM Tools" 2010 [Online]. Available: http://www.promtools.org/doku.php [Accessed: 04-09-2018]

[4] Fluxicon, "Disco: Discover your processes." [Online] Available: https://fluxicon.com/disco/ [Accessed: 04-09-2018]

[5] (Golfarelli and Rizzi 2009)
Golfarelli, Matteo and Stefano Rizzi. 2009. "*Data Warehouse Design,Modern Principles and Methodologies"*.

[6] (Kimball and Ross 2013)
Kimball, Ralph, and Margy Ross. 2013. "*The Data Warehouse Toolkit, The Definitive Guide to Dimensional Modeling*."

[7] (OMG 2010)
OMG. 2010. "*Business Process Model and Notation (BPMN)*."

[8] (Weske 2007)
Weske, Mathias. 2007. "*Business Process Management: Concepts, Languages, Architectures*."

[9] "Visual Paradigm"[Online]. Available: https://www.visual-paradigm.com/. [Accessed: 04-09-2018]

[10] Daley,Richard "Pentaho"[Online]. Available: https://www.hitachivantara.com/go/pentaho.html. [Accessed: 04-09-2018].

[11] Pentaho, "Mondrian | Pentaho Community." [Online]. Available: http://community.pentaho.com/projects/mondrian/ . [Accessed: 04-09-2018].

[12] C. Guyer and O. Duncan, "Querying Multidimensional Data with MDX | Microsoft Docs," 2017. [Online]. Available: https://docs.microsoft.com/en-us/sql/analysis-services/multidimensional-models/mdx/querying-multidimensional-data-with-mdx?view=sql-server-2017. [Accessed:04-09-2018].

[13] S. Wood, "Pentaho Mondrian Documentation," 2007. [Online]. Available: http://mondrian.pentaho.com/documentation/workbench.php . [Accessed: 04-09-2018].

[14] Pentaho, "Pentaho Data Integration. "[Online]. Available: https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html .[Accessed: 04-09-2018]

[15] TechTalk [Online]. Available: https://techtalk.gfi.com/the-top-20-free-network-monitoring-and-analysis-tools-for-sys-admins/ .[Accessed: 04-09-2018]

[16] Wikipédia [Online]. Available: https://pt.wikipedia.org/wiki/Sistema_de_Nomes_de_Dom%C3%ADnio . [Accessed: 04-09-2018]

[17] G. Kessler and S. Shepard "Network Working Group" [Online]. Available: https://tools.ietf.org/html/rfc2151#section-3.4 .[Accessed: 04-09-2018]

[18] Gaudenz Alder and David Benson "Draw.io" [Online]. Available: https://about.draw.io/ .[Accessed: 04-09-2018]

[19] bidashboard, [Online]. Available: https://www.bidashboard.org/. [Accessed: 04-09-2018]

[20] University of Zagreb "CORE network emulator" [Online]. Available: https://www.nrl.navy.mil/itd/ncs/products/core . [Accessed: 04-09-2018]

[21] techtalk [Online]. Available: https://techtalk.gfi.com/the-top-20-free-network-monitoring-and-analysis-tools-for-sys-admins/ .[Accessed:11-09-2018]

[22] selecthub [Online]. Available: https://selecthub.com/server-monitoring-software/ .[Accessed:11-09-2018]

[23] GFISoftware[Online]. Available: https://www.gfi.com/ . [Accessed:11-09-2018]

[24] HSPI Management Consulting "Process mining: A DATABASE OF APPLICATIONS" [Online]. Available: http://www.win.tue.nl/ieeetfpm/lib/exe/fetch.php?media=news:process_mining_database_applications_2017.pdf. [Accessed:11-09-2018]

[25] Nederlandse Spoorwegen [Online]. Available: https://www.ns.nl/en. [Accessed:11-09-2018]

[26]Fluxicon[Online].Available:https://fluxicon.com/blog/2018/05/process-mining-at-the-dutch-railway-process-mining-camp-2017/ . [Accessed:11-09-2018]

[27] Academic Medical Center [Online]. Available: https://www.amc.nl/web/over-de-locatie-amc/organisatie/about-the-amc.htm. [Accessed:11-09-2018]

[28] (Mans et al. 2008)
Mans, R. S., M. H. Schonenberg, M. Song, W. M.P. Van Der Aalst, and P. J.M. Bakker. "Application of Process Mining in Healthcare - A Case Study in a Dutch Hospital."

[29] Vodafone [Online]. Available: https://www.vodafone.pt/. [Accessed:11-09-2018]

[30] Celonis [Online]. Available: https://www.celonis.com/. [Accessed:11-09-2018]

[31] SAP [Online]. Available: https://www.sap.com/products/hana.html. [Accessed:11-09-2018]

[32] Celonis and Vodafone[Online]. Available: https://www.celonis.com/blog/vodafone-making-data-speak/. [Accessed:11-09-2018]

[33] Barros, Rui Miguel Pereira da Costa "Dashboarding - Projeto e Implementação de Painéis Analíticos."

[34] Oracle," Mysql" [Online]. Available: https://www.mysql.com/. [Accessed :11-09-2018]

[35] Quadient, "Data Cleaner" [Online]. Available: https://datacleaner.org/. [Accessed :11-09-2018]

Lista de Siglas e Acrónimos

ETL *Extract Transform Load*

CSV *Comma-separated values*

BPMN *Business Process Model and Notation*

IP *Internet Protocol*

DNS *Domain Name System*

XML *Extensible Markup Language*

OLAP *Online Analytical Processing*

KPI *Key Performance Indicator*

VPN *Virtual Private Network*

IoT *Internet of Things*

IBM *International Business Machines*

CPU *Central Processing Unit*

Anexos