Determinação da Qualidade de Execução de um Sistema de Povoamento de um Data Warehouse

Determining the Quality of Execution of a Data Warehousing Populating System

Nuno Dias, Centro ALGORITMI, Universidade do Minho, Portugal, a67749@alunos.uminho.pt

Orlando Belo, Centro ALGORITMI, Universidade do Minho, Portugal, obelo@di.uminho.pt

Resumo

Qualquer sistema de ETL (Extração, Transformação e Carregamento) que lide com um conjunto de dados significativo tem sempre a necessidade de contornar eventuais fenómenos de omissão e de inconsistência de dados que, frequentemente, resultam de uma ineficaz implementação dos sistemas operacionais associados. Caso se atue de forma contrária, o sistema poderá perder utilidade. Para que tais situações sejam raras, é preciso que, desde a fase embrionária do processo, se identifique, caracterize e solucione potenciais pontos de estrangulamento no desempenho do sistema de ETL. Neste trabalho reportamos um processo que desenvolvemos para determinar a qualidade da execução de um sistema de ETL, identificando e caracterizando eventuais pontos de estrangulamento – pontos negros – e a partir daí gerar um índice de qualidade de desempenho que nos apresente o "bem-estar" do sistema e nos forneça informação para a resolução de eventuais pontos negros e consequente melhoria da qualidade de serviço do sistema.

Palavras-chave: Sistemas de Data Warehousing; ETL; Process Mining; Otimização de Desempenho; Pontos Negros de ETL; Índices de Qualidade de Serviço.

Abstract

Any ETL (Extract, Transform and Load) system that deals with a meaningful dataset will always need to circumvent any phenomena of data omission and inconsistency. In most cases this results from an ineffective implementation of the associated operational systems. If the ETL system acts contrarily, the system may lose utility. For these situations being a rarity, it is necessary, from an embryonic stage of the process, to identify, characterize and solve potential bottlenecks in system execution. In this work, we report the process that we developed to determine the quality of the execution of an ETL system, identifying and characterizing eventual bottlenecks - black spots - and from that point on, generate a quality index of performance that presents us with the "well-being" of the system and provides us with information for solving any black spots and consequently improving the system's quality of service.

Keywords: *Data Warehousing Systems; ETL; Process Mining; Performance Optimization; ETL Black Points; Quality of Service Indexes.*

1. INTRODUÇÃO

Os sistemas de ETL são entidades fundamentais na intermediação entre os sistemas operacionais e os sistemas de *data warehousing* (Vassiliadis & Simitsis, 2009). De facto, são eles os processos responsáveis pela extração de dados dos sistemas operacionais, pelo seu tratamento para resolução

ISSN 2183-489X
DOI http://dx.doi.org/10.18803/capsi.v17.00-00

de possíveis inconsistências, omissões e outro tipo de problemas, e pelo seu carregamento na estrutura final do *data warehouse*, i.e., nas suas tabelas de dimensão e de factos. Daqui já se conseguem retirar algumas conclusões quanto à relevância e pertinência deste tipo de sistemas num sistema de *data warehousing*. Porém a sua importância não surge dissociada da sua dificuldade de implementação, sobretudo devido à quantidade de tarefas a que um processo ETL deverá dar resposta (Gour, Sarangdevot, Tanwar, & Sharma, 2010). Estes processos são muitas vezes implementados seguindo uma metodologia *ad-hoc*, despreocupada, apesar de envolver questões muito relevantes como o desempenho do sistema, a correção dos seus processos de conciliação de dados ou o tratamento de situações que possam envolver disparidade de formatos de dados ou a existência de valores imprevistos. Nesse sentido, quando um processo de ETL parece terminado, substituindo-se nessa altura os dados exemplo por dados reais, o cenário aplicacional costuma frequentemente complicar-se com o surgimento de inesperadas situações anómalas que revelam erros de implementação ou elementos de dados que, não tendo sido devidamente filtrados e tratados, passaram à etapa final de carregamento do *Data Warehouse* alterando o seu estado de forma indesejável. A isto, podemos também acrescentar que, quando não são os erros *de* implementação a travar a execução do processo, costumam emergir várias questões relacionadas com o desempenho do sistema e com a sua qualidade de serviço. Tudo isto faz com que seja necessário introduzir alguma linearidade no processo de implementação, bem como mecanismos que permitam, a cada momento, determinar a qualidade da execução do sistema de ETL.

A verificação da qualidade de execução dos sistemas ETL é, portanto, uma temática de extrema relevância ao permitir verificar potenciais pontos de quebra no seu desempenho e, assim, traçar ao longo do tempo, um perfil de execução do processo. Essa qualidade de execução pode ser medida através da definição de um índice de bem-estar, cuja variação permitirá verificar rapidamente a forma como o sistema se está a comportar e identificar eventuais situações anómalas. Se suportarmos a geração e manutenção deste índice de bem-estar com informação de suporte proveniente de uma base de dados multidimensional especialmente desenhada para o efeito e orientada por requisitos de *process mining*, podemos obter uma indicação precisa sobre a execução do sistema, com base em informação relacionada com os eventos realizados ao longo do tempo no sistema, o volume de dados envolvido, ou a janela de oportunidade disponível, entre outros.

Neste artigo apresentamos a solução que implementámos para a manutenção e caracterização de um índice de qualidade de serviço para um sistema de ETL, discutindo a forma como através da utilização de técnicas de mineração de processos identificamos os pontos de estrangulamento do sistema e geramos a informação necessária para alimentar o processo de manutenção do índice de bem-estar dos sistemas. Assim, de seguida, fazemos uma breve exposição de alguns trabalhos relacionados com este que aqui expomos, revelamos a forma como identificamos os pontos de estrangulamento e definimos um perfil de execução para um sistema de ETL, e geramos os valores

para alimentar o índice de bem-estar e acompanhar a sua variação ao longo do tempo. Por fim apresentamos algumas conclusões sobre o trabalho realizado e apontamos algumas linhas para desenvolvimento futuro.

2. TRABALHO RELACIONADO

A aplicação de técnicas de *process mining* em organizações é cada vez mais frequente em meios empresariais. Apesar disso, nem todas as técnicas se revelam adequadas em determinados contextos dadas as limitações que apresentam no tratamento de alguns casos específicos de aplicação. No caso particular da aplicação de técnicas de *process mining*, a obtenção de *logs* de eventos dos processos que queremos tratar é uma necessidade fundamental. Contudo, nem sempre são fáceis de obter, mesmo quando a disponibilidade de acesso aos registos do processo não se revela um problema. O nível de detalhe das *logs* de eventos deverá ser avaliado no sentido de se verificar se é (ou não) o mais adequado para se reconstituir o processo que lhe deu origem. Além disso, não podemos descurar, também, a forma como a própria *log* de eventos está definida, uma vez que frequentemente os eventos estão mal definidos ou registados em formatos incompatíveis com os processos de tratamento e análise. As ferramentas de geração automática de *logs* de eventos ajudam um pouco a ultrapassar este tipo de situações indesejáveis, uma vez que facilitam bastante o processo de obtenção das *logs* de eventos. No entanto, muitas destas ferramentas não registam os eventos de uma forma explícita, o que condiciona a sua aplicação (van der Aalst, 2015). Os modelos descobertos a partir de técnicas de mineração de processos não deverão ser encarados como representações totalmente fidedignas da realidade que se pretende modelar. Nesse sentido, a aplicação de técnicas de avaliação de um modelo descoberto para analisar *bottlenecks*, identificar ineficiências processuais, verificar a conformidade de dados, explicar desvios comportamentais, prever desempenhos e orientar os utilizadores na implementação de melhores processos não deverá ser descurada. Na realidade deve-se sempre pensar num cenário tipicamente iterativo, desde o primeiro modelo descoberto até ao modelo dito "final", no caso de se pretender melhorar os resultados da aplicação de um qualquer sistema de *process mining* (van der Aalst, 2015). A escolha de uma ferramenta de aplicação de *process mining* a uma *log* de eventos não é um processo simples. Pelo contrário, uma vez que condiciona, irremediavelmente, a qualidade do modelo produzido. Além disso, a instalação e manutenção das infraestruturas que asseguram a "ponte" entre os sistemas de informação e as ferramentas de mineração, além de estar longe de ser trivial, revela-se frequentemente bastante custosa. Com efeito, o mapeamento das informações recolhidas na ferramenta é, tão só, um exemplo de tarefa crucial no âmbito da aplicação de *process mining* (B.F. van Dongen, 2005).

Os processos não estruturados representam, também, uma dificuldade acrescida para as ditas tradicionais técnicas de mineração de processos, tornando-se importante fazer a distinção entre os

casos principais e os casos desviantes, para que estes sejam separados e não condicionem a descoberta de um modelo de mineração. Caso tal não aconteça, esses casos poderão estar a integrar o modelo final e a contribuir para resultados potencialmente enganosos (Hompes, Buijs, van der Aalst, Dixit, & Buurman, 2015). Contudo, os ambientes mais flexíveis são aqueles que são os mais atrativos para aplicação de técnicas de *process mining*. Porém, como a dificuldade de análise é proporcional à complexidade do processo a montante, neste tipo de casos, costuma-se optar por separar os dados em vários *clusters*, cada um contendo subconjuntos de dados homogéneos, e gerar para cada um deles um modelo de processo específico. Como tal, os resultados em ambientes flexíveis são tendencialmente melhores (Song, Günther, & Van Der Aalst, 2009).

A área da mineração de dados e suas técnicas não está circunscrita ao *process mining*. Com efeito, e dependendo do problema a tratar, perfilam-se no domínio muitas outras técnicas, como é o caso do *pattern mining*, que incide na descoberta de padrões estatisticamente relevantes em processos cujos eventos ocorrem de forma parcialmente ordenada. Esta área de mineração de dados não tem em conta as instâncias dos processos. Como tal, o que se pretende é apresentar uma nova técnica, e respetiva representação, que consiga extrair episódios frequentes a partir de uma *log* de eventos, ao mesmo tempo que se tira partido da associação dos eventos com casos concretos. Mais do que descobrir padrões em processos complexos e avaliar a sua conformidade, tendo em conta a ordem de ocorrência dos eventos, através deste tipo de técnicas é possível descobrir, também, regras de previsão de comportamento (e de comportamentos correlacionados), aplicando as técnicas selecionadas a outras perspetivas que possam figurar nas *logs* de eventos. Isto, poderá ser, também, um bom ponto de partida para análise de pontos de estrangulamento na execução de processos ETL (Mannila, Toivonen, & Verkamo, 1997).

Em suma, a vastidão da área de *process mining* e respetivas técnicas não está, ainda, associada à sua eficácia e adequação face aos processos que pretende modelar. No entanto, entre a comunidade de investigadores, a necessidade de avaliar as técnicas de *process mining* tem vindo gradualmente a crescer, isto porque não é possível, atualmente, verificar ainda a qualidade dos modelos descobertos pelos algoritmos disponíveis, nem tão pouco avaliá-los ou compará-los em termos de desempenho. Contudo, já existem propostas para uma *framework* capaz de suportar tais pretensões (Rozinat, De Medeiros, Günther, Weijters, & Van Der Aalst, 2008).

3. IDENTIFICAÇÃO DE PONTOS DE ESTRANGULAMENTO DE EXECUÇÃO

As ferramentas utilizadas no desenvolvimento de um sistema ETL condicionam irremediavelmente a qualidade das *logs* de eventos obtidas, quer em termos da sua disponibilidade, quer em termos da sua qualidade. Como tal, é necessário definir mecanismos alternativos para que essas *logs* sejam suficientemente detalhadas (e completas) para que se possa realizar um processo de mineração com sucesso. Antes de se realizar qualquer processo de monitorização de um sistema de ETL, é

preciso saber, em concreto, aquilo que pretendemos monitorizar, isto é analisar a forma como as tarefas escalonadas foram, de facto, realizadas e aquilo que envolveram na sua execução. Com base no trabalho realizado por Torres et al. (2016) selecionou-se um processo ETL específico, com as características adequadas para o tipo de trabalho de rastreio que se pretendia realizar. O sistema selecionado é responsável pelo povoamento de um *data warehouse* que acolhe a informação de suporte à decisão relacionada com as atividades de negócio de um pequeno *cluster* de três empresas. De referir que, o sistema de ETL em questão foi analisado anteriormente com o objetivo de fazer a identificação dos seus possíveis pontos de estrangulamento de execução – pontos negros - ao longo dos seus vários ciclos de execução (Belo, Dias, Pinto, & Ferreira, 2017). Para isso, foi necessário fazer o enriquecimento da sua *log* de eventos diária, introduzindo vários elementos de dados pertinentes, como sejam os casos de *time stamping* associados ao início e término de cada tarefa e à respetiva análise da rede de execução que seria posteriormente gerada numa ferramenta de mineração de processos - *Disco* (Disco, 2014). Dadas as características específicas de cada uma das empresas do *cluster*, o sistema ETL precisa de ter em conta vários fatores fundamentais, o primeiro dos quais é bastante crítico: a janela de oportunidade, o período de tempo que o sistema tem para realizar o seu trabalho. Além disso, o sistema de ETL tem que realizar várias operações de extração de dados, tarefas usualmente bastante complexas, para angariar os dados requeridos pelos sistema de *data warehousing* nos vários sistemas operacionais (duas fontes de dados são bases de dados relacionais e a terceira produz uma folha de cálculo com os dados de atividade diários), que requerem processos de extração, limpeza e conciliação bastante diversificados e exigentes, tanto em termos de elementos de dados como de execução de tarefas. O processo de ETL é executado, diariamente, entre as 6h e as 8h (fuso horário português) por invocação de um *package* implementado em *Kettle* (Pentaho, 2017) .

Caracterizado que está o sistema de ETL, vejamos agora a forma como estabelecemos o processo para a determinação da qualidade de execução do sistema de ETL. A qualidade de execução pode-se medir através do desempenho do processo, uma vez que este permite avaliar a forma como executa as tarefas definidas no menor intervalo de tempo possível. Assim, a identificação de pontos de estrangulamento é, de certa forma, um processo de avaliação da linearidade e qualidade do processo ETL. No entanto, nem todos os pontos de execução que revelem um tempo de execução anómalo são na realidade pontos negros. Nesse sentido, antes de se tecer qualquer consideração acerca da classificação de um ponto de execução como um ponto negro é necessário avaliar o seu impacto no processo como um todo e definir um *threshold* mínimo a partir do qual se possa considerar que um ponto de execução é um ponto negro. A definição do *threshold* deverá ter em consideração as características do próprio processo de ETL em análise.

4. GERAÇÃO DO PERFIL DE EXECUÇÃO DO SISTEMA ETL

Uma vez estabelecido e validado o processo de descoberta de pontos negros, e no sentido de determinar a qualidade do processo de execução do sistema ETL, ao longo dos seus vários ciclos de execução, é necessário definir e implementar uma infraestrutura que seja adequada ao tipo de análise pretendido. Com efeito, para caracterizar cada ponto negro em particular, bem como as circunstâncias subjacentes ao seu aparecimento, decidiu-se implementar um sistema de dados multidimensional, cuja estrutura foi orientada especificamente pelos requisitos mais relevantes do processo de análise a implementar. Para que fosse possível passar à sua implementação, foi necessário extrair as redes do processo para um formato que fosse facilmente manipulável - .csv.

Os dados recolhidos foram enriquecidos com a data e hora de execução das várias tarefas executadas pelo processo de ETL e cada passo identificado com a fase do processo no qual foi utilizado, de forma a que fosse possível realizar uma análise mais aprofundada e traçar um perfil temporal dos ciclos de execução do sistema de ETL – são elementos fundamentais no sustento do sistema de análise. Porém, antes de se realizar um primeiro esboço da estrutura do sistema multidimensional, foi necessário definir e caracterizar o grão para a estrutura de dados de suporte, ou seja, o nível mais elementar da informação que iria ser armazenada no sistema multidimensional relacionada com a avaliação da execução do sistema de ETL. Assim, definiu-se o grão como sendo a peça de dados mais elementar com capacidade para caracterizar um ponto negro, relativo a um determinado ciclo de execução do processo ETL. Além disso, definiram-se também as diversas perspetivas de análise a realizar sobre os pontos negros que deram origem ao seguinte conjunto de tabelas de dimensão: *Calendário*, para suportar a análise dos pontos negros segundo um dado ponto de vista temporal; *CaseGrupo*, para verificar a(s) transformação(ões) em que foi utilizado um dado componente *Kettle*; e *Componente*, para identificar quais os componentes *Kettle* envolvidos na génese de um determinado ponto negro.

Para além destes elementos, o esquema dimensional do sistema (Figura 1) incorpora uma única tabela de factos (*FT_PontoNegro*), que integra como medidas de análise, nomeadamente, a frequência absoluta de utilização (*freqAbs*), o número de transformações em que aparece (*freqCase*), o número máximo de vezes em que foi utilizado numa transformação (*caseFreq*), a duração total (*duracaoTotal*), a mediana da duração (*duracaoMediana*), a duração média (*duracaoMedia*), a duração mínima (*duracaoMin*) e a duração máxima (*duracaoMax*). Complementarmente, acrescentaram-se algumas medidas de controlo, entre as quais se destacam as medidas *bemEstar_parcial* e *isBlackpoint*. A primeira destas medidas diz respeito ao cálculo do índice de bem-estar de forma parcial, i.e., para apenas um componente *Kettle*[1], enquanto que a segunda medida, *isBlackpoint*, permite identificar os casos que são pontos negros

[1] Verificar secção 5, equação 1.

(*isBlackpoint*=1), i.e., os casos cuja mediana do tempo foi superior ou igual ao *threshold* definido para o ETL. Ambas as medidas integram o cálculo do índice de bem-estar final do sistema ETL.

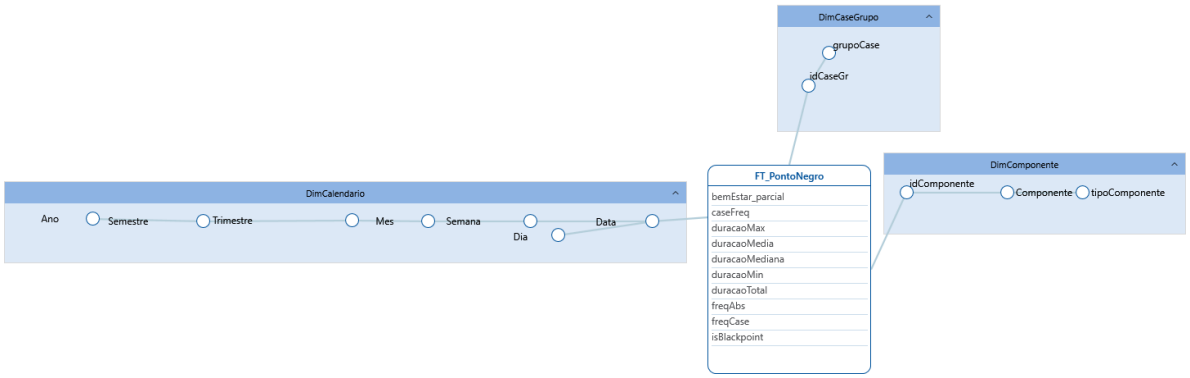

Figura 1 - Esquema dimensional relativo ao *data warehouse BlackpointDW*

Uma questão fundamental a considerar, quando se analisam as métricas resultantes do processo de mineração, é a influência de casos de exceção (*outliers*) no processo de deteção de pontos negros. Na definição de *outlier* incluem-se todas as medições de tempo anómalas que fazem parte do cálculo do tempo médio de processamento de cada passo do processo de ETL. Ao fazerem parte da expressão de cálculo, estas medições poderão estar a contribuir para uma identificação errada de pontos negros, o que pode conduzir a uma hipotética sobrevalorização de um determinado ponto de execução em detrimento de outros.

Assim, e tendo em conta o facto da ferramenta *Disco* calcular sempre a média e a mediana do tempo, bem como a natureza do processo que está a ser analisado, a mediana é, no nosso ponto de vista, aquela que permite fazer uma melhor aproximação ao tempo que cada passo ETL demora a executar, uma vez que, pela sua própria fórmula de cálculo, exclui valores anómalos, i.e., valores demasiado altos ou valores demasiado baixos (Gunther & Rozinat, 2014).

Vejamos, agora, as diferenças das imagens apresentadas na Figura 2 e na Figura 3. Como facilmente se percebe a utilização do tempo médio, como filtro de análise, ao invés da mediana do tempo, identifica pontos negros que na realidade não passam de "falsos" pontos negros (a cor mais escura), uma vez que estes são influenciados por um valor anormalmente elevado entre os vários valores recolhidos ao longo dos ciclos de execução do sistema ETL. Isto fez com que se optasse por utilizar o tempo mediano de execução de um passo, em substituição do tempo médio.

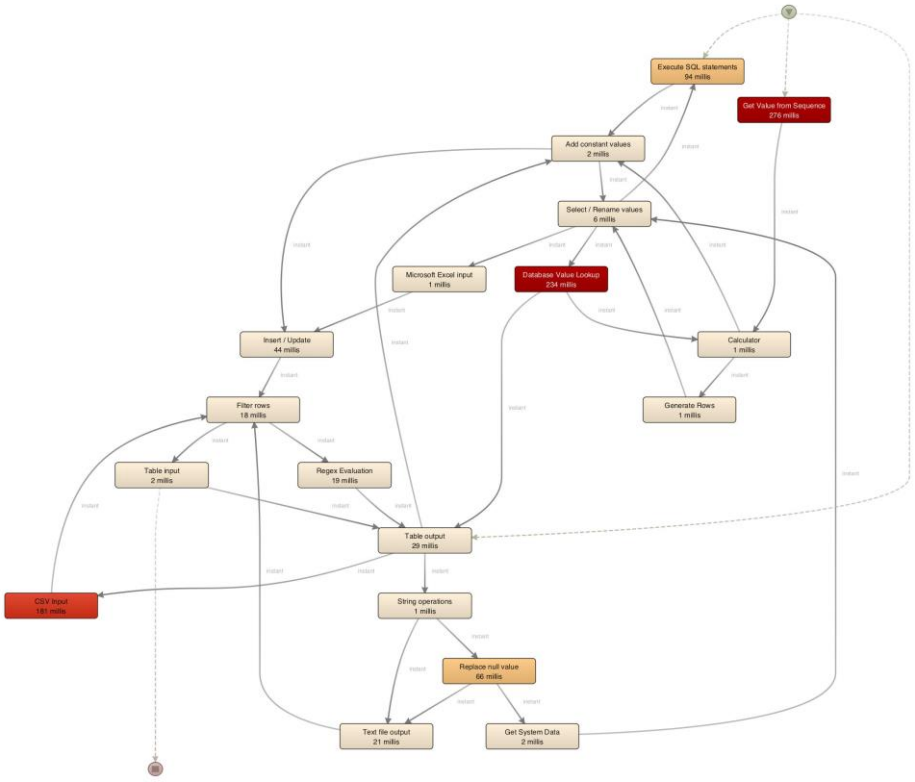

Figura 2 – Exemplo de uma rede de execução com filtro - tempo médio.

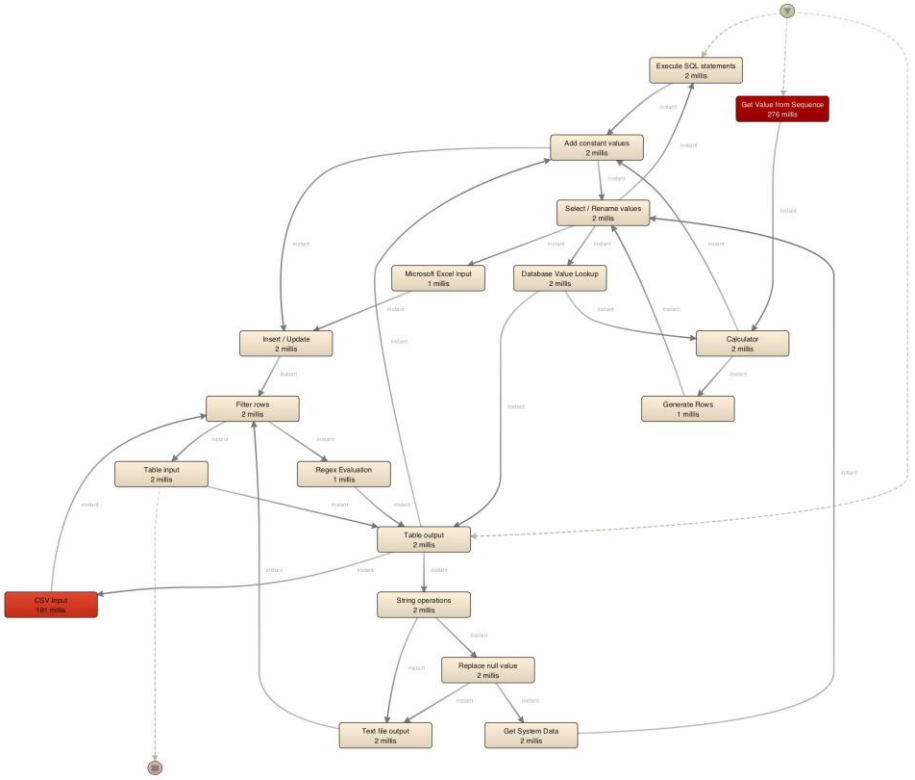

Figura 3 - Exemplo de uma rede de execução com filtro - tempo mediano.

5. O ÍNDICE DE BEM-ESTAR ETL

O índice de bem-estar ETL permite determinar a qualidade de execução do sistema de ETL ao longo dos vários ciclos de execução do sistema de povoamento. Com efeito, o facto deste índice ser calculado com base nos resultados da aplicação de *process mining* às suas *logs* de eventos permite que se perceba a origem de uma situação anómala de uma forma expedita, com recurso ao sistema de dados multidimensional criado para o efeito. Porém, a forma de fazer o cálculo deste índice não foi um processo fácil. Primeiramente, foi necessário definir quais as métricas resultantes do processo de mineração que iriam integrar a expressão que suportaria o seu cálculo, bem como as métricas que deveriam ser obtidas de forma indireta ao processo. Posteriormente, definimos a forma como todas estas métricas se iriam relacionar de forma a estabelecerem a base de cálculo do índice de bem-estar do sistema de ETL. A lista de métricas exportadas da rede do processo gerada pela ferramenta *Disco* inclui, para cada componente *Kettle*, a frequência absoluta, a frequência de *case*, o máximo de repetições, a duração total, a duração mediana, a duração média, a duração mínima e a duração máxima. Para posterior entendimento do porquê da inclusão ou exclusão de uma métrica, e tendo em conta o contexto do problema em questão, vejamos com um pouco mais de atenção a definição de cada uma das medidas consideradas:

- Frequência absoluta – o número total de vezes que um componente *Kettle* foi utilizado num *package* ETL.
- Frequência de *case* – o número de transformações (*cases*) distintas em que um componente *Kettle* é utilizado.
- Máximo de repetições – o número máximo de vezes que um componente *Kettle* é utilizado numa só transformação.
- Duração total – o tempo total, em ms, que um determinado componente ocupou na execução de um *package* ETL.
- Duração mediana – a mediana do tempo, em ms.
- Duração média – a média do tempo, em ms, que um determinado componente *Kettle* ocupa relativamente ao tempo total de execução do *package*.
- Duração mínima – o tempo mínimo (em ms) que um determinado componente *Kettle* demorou a ser executado numa transformação.
- Duração máxima – o tempo máximo (em ms) que um determinado componente *Kettle* demorou a ser executado numa transformação.

Activity	Absolute frequency	Case frequency	Max. repetitions	Total duration (ms)	Median duration (ms)	Mean duration (ms)	Min. duration (ms)	Max. duration (ms)
Execute SQL statements	28	10	4	2646	2	94	0	1863
Select / Rename values	46	13	6	318	2	6	1	206
Filter rows	48	14	7	886	2	18	1	287
Table output	46	18	5	1369	2	29	1	268

Figura 4 - Métricas resultantes do processo de mineração de dados pela ferramenta *Disco*.

Das métricas referidas, foram selecionadas a frequência absoluta, a frequência de *case* e a duração mediana para o processo de análise. As restantes medidas são, neste caso, irrelevantes para o cálculo do índice de bem-estar.

Para o índice de bem-estar de um processo ETL, fatores como o período temporal de execução são relevantes. O sistema ETL que sustenta o sistema multidimensional em estudo, ao estar ligado à área comercial torna possível associar o fator tempo ao volume de dados observado ao longo dos vários ciclos de execução. Assim sendo, essa periodicidade deveria ser considerada de alguma forma na expressão final. Pensou-se, então, em incluir o volume de dados, mais concretamente o número de registos associados a operações de *insert*, *update* e *delete*, sabendo à partida que o número de novos registos seria, na maioria dos casos, superior ao número de registos marcados como *update* ou *delete*. Convém realçar, de igual forma, que os sistemas ETL implementam, por norma, políticas bem definidas no que diz respeito a operações de *update* e *delete*, sendo que, não raras vezes, certas dimensões nem consideram tais tipos de operação. Por outro lado, o impacto da fase do processo ETL na qual um determinado componente *Kettle* é utilizado foi, de igual forma, considerado. Para isso calculou-se a proporção (peso) dos vários componentes utilizados na fase de extração, limpeza, transformação e carregamento, através do cálculo da razão entre a frequência absoluta de utilização e o somatório das frequências absolutas. Além disso, o consumo de energia de um *package* ETL é uma métrica bastante interessante a ter em conta no atual contexto de crescente preocupação ambiental, no qual a poupança de energia é praticamente um imperativo. Assim sendo, e com base no trabalho realizado por (Guimarães, Saraiva, & Belo, 2015), integrou-se, também, esta métrica no processo de cálculo do índice de bem-estar do sistema de ETL (Equação 4). Outro fator muito relevante para qualquer sistema ETL é a janela de oportunidade. Assim, a proporção de janela de oportunidade que, em cada ciclo de execução, é ocupada (um período de tempo) será uma outra métrica a incluir no processo de cálculo do índice de bem-estar. Para concluir, falta apenas ter em conta o valor de *threshold*, i.e., um valor mínimo de tempo a partir do qual se considera que um determinado componente *Kettle* é efetivamente um ponto negro, e que deverá ser ajustado de acordo com a natureza do processo ETL em análise.

A expressão de cálculo para o índice de bem-estar apresenta-se de seguida (Equação 1). De notar que o somatório que figura nessa equação depende de uma restrição e que o valor de $f(x)$ depende do valor que, em cada momento, a expressão $\frac{t_{total} \times 2}{janela_{oport.}}$ possa assumir (Equação 2).

$$índice_{bem-estar} = \frac{\#insert}{\#insert + \#update + \#delete} \times f\left(\frac{t_{total} \times 2}{janela_{oport.}}\right) \times \sum_{dur_{mediana} \geq threshold} p.\left(\frac{f_{abs}}{case_{freq.}}\right).dur_{mediana} + E \quad (1)$$

$$f(x) = \begin{cases} 1 \; se \; x \leq 1 \\ \left(\frac{t_{total}}{janela_{oport.}}\right) \; se \; x > 1 \end{cases} \quad (2)$$

$$p_i = \frac{\#componentes \; etapa \; i}{\#componentes \; total}, \quad (3)$$

$$i \in \{extração, transformação, carregamento, limpeza\}$$

$$\mathrm{E} = \mathrm{P}.\Delta\mathrm{t}, \quad (4)$$

Quando a razão $\left(\frac{t_{total} \times 2}{janela_{oport.}}\right) \leq 1$, então assume-se $\left(\frac{t_{total}}{janela_{oport.}}\right) = 1$ pois o tempo de execução do *package* ETL não está a comprometer uma possível necessidade de *rollback* e repetição do processo por motivo de erro ou exceção (a janela de oportunidade continua a ser suficiente).

Para concluir, os factos guardados no sistema multidimensional de dados já têm o índice de bem-estar parcialmente calculado, mais precisamente o cálculo $p.\left(\frac{f_{abs}}{case_{freq.}}\right).dur_{mediana}$. O índice de bem-estar tenderá a aumentar, quanto pior for o desempenho do sistema de ETL. Repare-se que, ao longo dos vários ciclos de execução, a razão entre a frequência absoluta de utilização de um componente e o número de *cases* em que esse componente é utilizado, se mantém praticamente constante, sendo sobretudo a duração mediana aquela que sofrerá variações mais consideráveis. Caso essa duração aumente, o valor do índice aumenta, o que nesse caso significa uma degradação na performance global do ETL. Em termos de variação, o índice de bem-estar variará sempre no intervalo [0,1]. Para melhorar a leitura e análise do índice de bem-estar em formato gráfico, aplica-se uma expressão matemática de normalização, assegurando que o valor do índice variará sempre no intervalo [0,1].

$$g(x) = \frac{x}{x^2+1} \quad (5)$$

No caso do ETL utilizado como caso de estudo, calculou-se para oito dias de execução o índice de bem-estar. De seguida povoou-se o *data warehouse* implementado com esses mesmos dados, construindo-se a estrutura multidimensional de suporte e associando-se essa estrutura a uma ferramenta de *dashboarding*. Terminado esse processo, obteve-se o gráfico de variação temporal do índice de bem-estar, por manipulação do cubo com interrogações MDX (Figura 5).

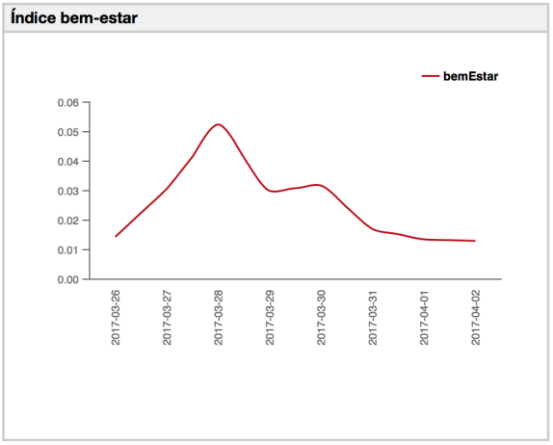

Figura 5 – Variação do índice de bem-estar do ETL ao longo do período de análise.

Por observação da Figura 5, rapidamente se percebe a utilidade deste indicador quando ilustrado com recurso a ferramentas de *dashboarding*. Convém relembrar que o gráfico apresentado resulta da aplicação da expressão de normalização (Equação 5) aos resultados obtidos, pelo que valores elevados de índice de bem-estar normalizado correspondem a valores baixos do índice de bem-estar, ou seja, situações de normalidade de performance do ETL. Como já era expectável o primeiro dia corresponde a uma situação atípica no desempenho do sistema, uma vez que o volume de dados a extrair é tendencialmente superior ao que sucede nos restantes dias. Assim sendo, o índice no gráfico atinge o seu valor mínimo no intervalo nesse momento, uma vez que estamos perante uma situação exigente em termos de desempenho.

6. CONCLUSÕES E TRABALHO FUTURO

Um sistema ETL é uma peça de *software* usualmente complexa. Na sua arquitetura e implementação muitos aspetos são considerados, envolvendo questões relacionadas direta ou indiretamente com o tipo de componentes a utilizar, operações a realizar, tipo e volume de dados envolvido, desempenho do sistema, entre outros. Porém, neste trabalho, abordamos apenas um desses aspetos: o desempenho do sistema. Com efeito, neste artigo abandonámos um pouco a ideia de que um sistema de ETL deve ser visto como uma espécie de caixa-negra, algo que é veiculado frequentemente aquando da utilização destes sistemas. Assim, procurámos relevar a importância da análise das *logs* de execução dos sistemas de ETL, como um instrumento essencial na verificação de potenciais anomalias que ocorram durante a execução de um sistema de ETL, quer estas emirjam de simples elementos estruturais ou de processos mal implementados. Mais do que aumentar o conhecimento de quem implementou o processo ETL, este artigo sublinha a utilidade de realizar essa avaliação em permanência, através da implementação de um índice de "bem-estar" que evidencie o nível da qualidade de serviço que o sistema está a prestar num dado momento. A utilidade deste índice só se torna mais evidente pelo papel crítico que estas infraestruturas

desempenham nas mais diversas organizações em termos de auxílio no processo de tomada de decisão. Por último, interessa também referir a importância da existência de um índice de bem-estar para um sistema de ETL, para o tempo presente (e passado), e tentar recorrer a uma camada de previsão para antecipar valores futuros para este tipo de índice e, desta forma, minimizar hipotéticos períodos de *downtime* nestes sistemas. Esta questão, deverá ser abordada futuramente, numa das linhas de desenvolvimento que temos planeadas para desenvolver a curto prazo.

REFERÊNCIAS

B.F. van Dongen, W. M. P. van der A. (2005). A Meta Model for Process Mining Data. *Proceedings of the CAiSE'05 Workshops*, *11*(i), 309–320.

Belo, O., Dias, N., Pinto, F., & Ferreira, C. (2017). Discovering ETL Black Points A Process Mining Approach. *5th World Conference on Information Systems and Technologies (WorldCIST 2017), Porto Santo Island, Madeira, Portugal, April, 11-13*.

Disco. (2014). Disco 1.6.0 — Flux Capacitor Update. Retrieved December 30, 2016, from https://fluxicon.com/blog/2014/01/disco-1-6-0/

Gour, V., Sarangdevot, S. S., Tanwar, G. S., & Sharma, A. (2010). Improve Performance of Extract, Transform and Load ({ETL)} in Data Warehouse. *International Journal on Computer Science & Engineering*, *1*(3), 786–789.

Guimarães, M., Saraiva, J., & Belo, O. (2015). Categorização do Consumo de Energia em Sistemas de Povoamento de Data Warehouses. *5º Conferência Da Associação Portuguesa de Sistemas de Informação (CAPSI'2015), Lisboa, Portugal, October, 2-3*.

Gunther, C. W., & Rozinat, A. (2014). Disco 1.6.0 — Flux Capacitor Median Performance Metrics. Retrieved February 13, 2017, from https://fluxicon.com/blog/2014/01/disco-1-6-0/

Hompes, B. F. A., Buijs, J., van der Aalst, W. M. P., Dixit, P. M., & Buurman, J. (2015). Discovering Deviating Cases and Process Variants Using Trace Clustering. *Proceedings of the 27th Benelux Conference on Artificial Intelligence (BNAIC), November*, 5–6.

Mannila, H., Toivonen, H., & Verkamo, A. I. (1997). Discovery of Frequent Episodes in Event Logs. *Data Mining and Knowledge Discovery*, *1*(3). http://doi.org/10.1023/A:1009748302351

Pentaho. (2017). Data Integration - Kettle. Retrieved from http://www.pentaho.com/product/data-integration

Rozinat, A., De Medeiros, A. K. A., Günther, C. W., Weijters, A. J. M. M., & Van Der Aalst, W. M. P. (2008). *The need for a process mining evaluation framework in research and practice: Position paper*. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4928 LNCS). http://doi.org/10.1007/978-3-540-78238-4_10

Song, M., Günther, C. W., & Van Der Aalst, W. M. P. (2009). Trace clustering in process mining. In *Lecture Notes in Business Information Processing* (Vol. 17 LNBIP, pp. 109–120). http://doi.org/10.1007/978-3-642-00328-8_11

Torres, B., Ferreira, C., Pinto, F., & Dias, N. (2016). *Um Data Warehouse para um cluster de empresas - Relatório Técnico Mestrado Integrado em Engenharia Informática Universidade do Minho*. Braga, Portugal.

van der Aalst, W. M. P. (2015). Extracting Event Data from Databases to Unleash Process Mining. *BPM - Driving Innovation in a Digital World SE - 8*, 105–128. http://doi.org/10.1007/978-3-319-14430-6_8

Vassiliadis, P., & Simitsis, A. (2009). Extraction, transformation, and loading. *Encyclopedia of Database Systems*, 10. http://doi.org/10.4018/987-1-59904-364-7.ch004