Extração de informação de ficheiros de log

Utilizando programação Python e a Ferramenta Tableau

Extraction of information from log files

Using Python Programming and Tableau

Filipe Rigueira
Coimbra Business School | ISCAC, Polytechnic of Coimbra
Coimbra, Portugal
a2019118806@alumni.iscac.pt

Jorge Bernardino
Instituto Politécnico de Coimbra – ISEC
i2A – Instituto de Investigação Aplicada
Coimbra, Portugal
jorge@isec.pt

Isabel Pedrosa
Coimbra Business School | ISCAC, Polytechnic of Coimbra
Instituto Universitário de Lisboa (ISCTE-IUL) ISTAR-IUL, Portugal
ipedrosa@iscac.pt

Resumo — **Os servidores aplicacionais geram ficheiros de logs diários onde se regista uma parte significativa da sua atividade. Essa informação é registada sequencialmente no tempo mas mistura vários tipos de informação. A ausência de um padrão para a formatação do registo de dados e o respetivo volume, dificultam a extração da respetiva informação. A inexistência de trabalhos, especificamente no tratamento de ficheiros de *logs* de servidores *Service-Oriented Architecture* (SOA), não permitiu, nem a utilização de métricas de controlo, nem um conjunto de boas práticas que pudessem ser seguidas. Resulta deste trabalho uma descrição do processo que poderá servir de guia: na definição de uma estrutura de registo de *logs*; na construção de um processo de extração de dados; na definição de uma estrutura de dados de suporte da informação extraída; na definição de métricas de controlo; na definição de processos de análise e controlo dos dados extraídos. Atendendo ao tamanho dos ficheiros e à diversidade de tipos de dados existentes foi necessário utilizar programação Python para a extração e pré-tratamento de dados, Excel para o pré-tratamento de dados, Tableau para o tratamento estatístico e apresentação dos resultados.**

Palavras Chave – Ficheiros de Logs; Key Performance Indicators; KPIs, Serviços SOA; Python; Tableau.

Abstract — ***Application servers generate daily log files with a significant part of their activity. This information is recorded sequentially over time but mixes various types of information. The absence of a standard for formatting the data record and the respective volume, make it difficult to extract the corresponding information. The lack of work, specifically in the treatment of SOA server log files, did not allow the comparisson with pre-existing Key Performance Indicators (KPI) or a set of best practices that could be followed. This work results in a description of the process that can serve as a guide for: definition of a logging structure; construction of a data extraction process; definition of a data structure to support the extracted information; definition of control metrics; definition of analysis and control processes for the extracted data.. Given the size of the files and the diversity of types of information that existed, it was necessary to use Python programming for data extraction and pre-treatment, Excel for data pre-treatment, Tableau for statistical treatment and presentation of results.***

Keywords – Log files, Key Performance Indicators, KPIs, Python, Tableau.

I. INTRODUÇÃO

Os servidores aplicacionais são geradores de muita informação que fica registada em ficheiros de *logs*. Esta informação é registada em ficheiros de texto sem que tenham uma lógica de formatação bem definida. No mesmo ficheiro podem constar dados com diferentes tipos de informação e diferentes estruturas, o que dificulta a sua análise e processamento. Na literatura encontram-se trabalhos que exploram o tratamento de ficheiros de *logs* na área dos *web services* [1] [2] e dos servidores SAP mas são raros os trabalhos na área dos servidores Service-Oriented Architecture (SOA) [3]. Este trabalho pretende colmatar essa falta, descrevendo um caso prático de extração e tratamento de dados a partir de ficheiros de *logs* de um servidor SOA. Não existindo um padrão definido para o registo de *logs*, a abordagem utilizada neste trabalho não poderá ser generalizada, no entanto poderá servir de guia para: a definição de uma estrutura de registo de *logs*; a construção de um processo de extração de dados; a definição de uma estrutura de dados de suporte da informação extraída; a definição de métricas de controlo; a definição de processos de análise e controlo dos dados extraídos.

Este trabalho encontra-se organizado da seguinte forma. Na secção II é descrito o problema e metodologia, e as fases do processo de análise de *logs*. A secção III analisa o conteúdo dos ficheiros. Nas seções IV, V, VI temos a identificação informação relevante, extração de dados, identificação de *Key Performance Indicators*, tratamento estatístico dos dados e Resultados e Discussão, respetivamente. O artigo termina com as Conclusões.

II. DESCRIÇÃO DO PROBLEMA E METODOLOGIA

O sistema de informação (SI) comercial foi desenhado numa lógica de camadas e está suportado por um conjunto de servidores, tal como é possível verificar na Figura 1. Os utilizadores acedem ao sistema a partir de aplicações do tipo cliente-servidor (desenvolvidas em Delphi) disponibilizadas por um servidor aplicacional (Sistema Operativo baseado em Windows). As regras de negócio e o acesso aos dados são garantidos pela camada de serviços (desenvolvido em Java Script), numa filosofia do tipo *SOA*, em servidor Unix. Os dados estão centralizados numa base de dados (BD) única.

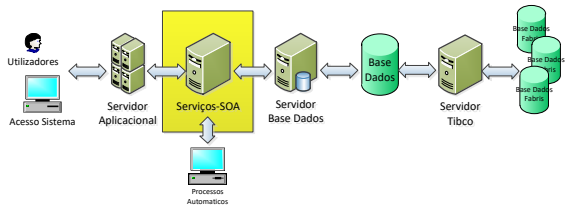


Figura 1 - Modelo do Sistema de Informação

O sistema é alimentado pelos utilizadores, por via das aplicações, e pelos sistemas dos diferentes polos fabris através de interfaces. Esta integração de dados entre sistemas é garantida por um servidor onde corre uma instância da ferramenta de integração que tem por função garantir a troca de dados entre as diferentes BD da organização.

O sistema, no seu conjunto, foi desenvolvido e implementado de forma a garantir a sua disponibilidade 24 horas por dia e 7 dias por semana.

A camada dos serviços SOA é uma parte fundamental de todo o sistema, é nessa camada onde residem e são validadas as regras de negócio.

O âmbito deste artigo está restrito à análise dos *logs* gerados pelos Serviços SOA e tem por objetivo: identificar toda a informação relevante do ponto de vista do SI, desenhar e implementar um processo de extração dessa informação, criar métricas que permitam efetuar um tratamento estatístico da informação recolhida, organizar a informação em tabelas e gráficos de forma a ter uma visão resumida da sua atividade.

Estes *logs* registam uma parte muito significativa da atividade do SI e contêm centenas de milhares de linhas organizadas sequencialmente no tempo. No mesmo ficheiro são registados eventos dos seguintes tipos de informação: *Exposing services* (disponibilização dos serviços e respetivos métodos); Logging Advice (chamadas aos serviços e respetivos métodos); Session Manager (registo de entrada e saída dos utilizadores no sistema); CCCException (erros e respetivas tipificações); DefaultFaultHandler (alertas e respetivas tipificações). Para cada um destes tipos de eventos a estrutura da informação é diferente sendo apenas comum a forma como é registado o momento do registo da sua ocorrência.

O volume de linhas, a diversidade de informação e a sequência temporal de registo, dificultam a respetiva leitura e análise pelo que esta informação é utilizada ocasionalmente na análise de erros.

Neste trabalho descrevem-se as etapas do processo: 1) Análise do conteúdo dos ficheiros; 2) Identificação da informação relevante; 3) Extração de dados; 4) Identificação de Key Performance Indicator (KPI); 5) Tratamento estatístico de dados; 6) Resultados; 7) Conclusão.

III. ANÁLISE DO CONTEÚDO DOS FICHEIROS

Da análise dos ficheiros resultam os seguintes tipos de informação:

- *Exposing Services*: registo da disponibilização dos serviços e respetivos métodos;
- *Logging Advice*: registo das chamadas aos serviços e respetivos métodos;
- *Session Manager*: registo dos acessos dos utilizadores ao Servidor SOA;
- *CCCException*: registo dos erros e respetivas tipificações;
- *DefaultFaultHandler*: registo de alertas funcionais e respetivas tipificações.

Todas as linhas começam com o registo do instante em que o evento ocorreu (YYYY-MM-DD hh:mm:ss,###) e um qualificador que identifica o tipo de informação: INFO/ERROR. A restante informação é registada de acordo com regra distinta para cada um dos respetivos tipos de informação.

IV. IDENTIFICAÇÃO DA INFORMAÇÃO RELEVANTE

Atendendo a que, para cada um dos tipos de linha, a informação registada é diferente, foi necessário, efetuar uma análise tipo a tipo, nomeadamente:

*Exposing Service*s – estas linhas correspondem à disponibilização dos serviços e respetivos métodos. Esta informação não foi considerada relevante, e como tal, foi excluída do processo de extração e tratamento de dados.

Session Manager - estas linhas contêm a informação dos registos de acessos dos utilizadores ao servidor SOA, pelo que se trata de informação relevante e de recolha e tratamento obrigatórios. Do respetivo conteúdo, foram identificados os seguintes dados: *Momento do registo*, *Tipo* (Create/Destroy), *session ID*, *userID*.

Logging Advice - estas linhas contêm a informação das chamadas aos serviços e respetivos métodos, pelo que se trata de informação muito relevante e necessária para uma visão global do tipo de utilização do SI. Foi identificada como sendo de recolha e tratamento obrigatórios. Do respetivo conteúdo, foram identificados os seguintes dados: *Momento do registo*, *userID*, *Serviço*, *Método*, *Número de parâmetros passados ao método*.

DefaultFaultHandler – estas linhas correspondem a respostas do serviço nos casos em que são identificadas excepções ou falhas de regras de negócio. Tratando-se de controlos funcionais, não foram considerados relevantes, pelo que foram descartados.

CCCException - estas linhas correspondem a erros pelo que o seu registo e tratamento são obrigatórios. Do respetivo conteúdo, foram identificados os seguintes dados: *Momento do registo*, *Descritivo do erro*, *userID*, *Serviço*, *Método*, linha do código onde ocorreu o erro.

```
2020-01-06 05:10:54,921 INFO [org.codehaus.xfire.spring.ServiceBean] Exposing service with name {ServicoID} ServicoID
…
2020-01-06 05:19:23,380 INFO [com.CS.CM.auth.entities.SessionManager] Session A1578287963380 created for user user_login
2020-01-06 05:19:05,202 INFO [com.CS.CM.auth.entities.SessionManager] Session A1578287963380 destroyed
…
2020-01-06 05:19:23,470 INFO  [LoggingAdvice]  [user_login]  [Serviço] metodo(true)
…
2020-01-03 15:54:40,716 ERROR [com.CS.CM.exception.CCCException] Descritivo do Erro.
User:   user_login
File:   ServiceID
Method: MetodoID
Line:   LineID
...
2020-01-03 15:48:58,465 INFO  [org.codehaus.xfire.handler.DefaultFaultHandler] Descritivo do Alerta
…
```

Figura 2 Exemplo de *logs*

V. EXTRACÇÃO DE DADOS

Como não foi possível utilizar a ferramenta normalmente utilizada para a extração de dados (Excel) foi necessário criar um programa para esse propósito (em Python). O programa desenvolvido percorre todas as linhas do ficheiro e, mediante a informação de cada linha, extrai apenas a informação anteriormente identificada como relevante, que é guardada em ficheiros separados por tipo (formato CSV) (Figura 3).

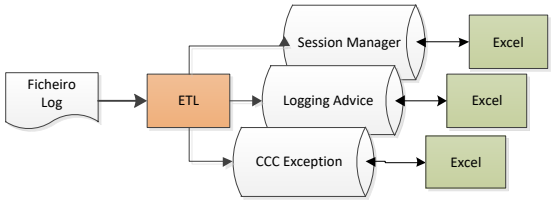


Figura 3 - Processo de extração de dados

Atendendo ao elevado número de linhas a processar, foi necessário dividir o processamento em blocos de informação o deu origem a linhas duplicadas. Para resolver este problema utilizou-se a funcionalidade do Excel que permite identificar e eliminar das tabelas as linhas duplicadas (ver Figura 4).

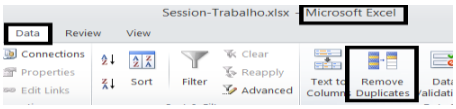


Figura 4 - Processo de extração de dados

O programa gera ficheiros distintos por tipo de informação com os seguintes conteúdos:

- *Session Manager*: (Ano, Mês, Dia, Hora, Minutos, Segundos, Selo-Inicio, Selo-Fim);
- *Logging Advice*: (Ano, Mês, Dia, Hora, Minutos, Segundos, Utilizador, Serviço, Método, Número de Parâmetros);
- *CCCException*: (Ano, Mês, Dia, Hora, Minutos, Segundos, Utilizador, Serviço, Método, Linha, Tipo Erro, Descritivo Erro). Os erros foram ainda classificados de acordo como Acessos a Dados/Erros Aplicacionais/Não Classificados.

VI. IDENTIFICAÇÃO DE KEY PERFORMANCE INDICATORS

A literatura disponível, relativamente ao tema, está muito direcionada para a utilização de técnicas de Data Mining na identificação de perfis de utilização qualitativa e não quantitativa [4]. Os trabalhos encontrados, quer seja no tratamento de *logs* de servidores web [1] [2], quer seja em *logs* de servidores SOA [3] focam-se na identificação de perfis de utilização funcional sem que sejam exista uma quantificação. Na impossibilidade de encontrar métricas de controlo de referência foram identificadas as seguintes:

Session Manager: N.º de Sessões Perdidas e Padrão Horário das sessões perdidas; N.º de sessões criadas, nº de utilizadores em sessões criadas, duração média das sessões c4iadas e Padrão horário das sessões criadas (Figura 5).

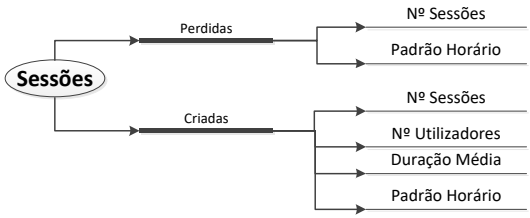


Figura 5 - Session Manager - Métricas

CCCException: padrão horário erros de acesso a dados, padrão horário de erros aplicacionais, Erros nos serviços de acesso a dados e erros nos serviços aplicacionais (Figura 6).

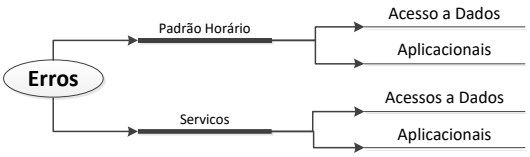


Figura 6 - CCCException - Métricas

Logging Advice: Padrão horário de chamadas a serviços de dados, Tipo de chamadas a serviços de dados, Padrão horário de chamadas a serviços (Figura7).

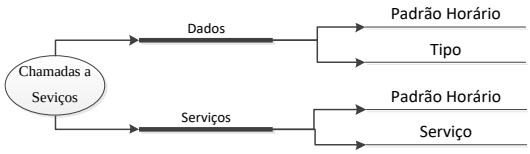


Figura 7 - Logging Advice - Métricas

VII. TRATAMENTO ESTATÍSTICO DE DADOS

Com base nos dados recolhidos e KPIs identificados foram desenvolvidos os seguintes cálculos, quanto à *Session Manager*: 1) Cálculo do tempo de ligação = Momento Final menos Momento Inicial mais 1; 2) Número de novas ligações: Agrupadas as ligações iniciadas durante a mesma hora; 3) Número de acessos de utilizadores: Contar o número de utilizadores que estabeleceram pelo menos uma ligação na mesma hora. Considerar cada utilizador apenas 1 vez independentemente do número de sessões por si estabelecidas; 4) Número de ligações perdidas: Contar o número de ligações perdidas na mesma hora; 5) Tempo médio de ligação (por Dia): Somar o tempo de todas as ligações e dividir pelo total de ligações. Considerar apenas ligações em que exista um registo de entrada e de saída.

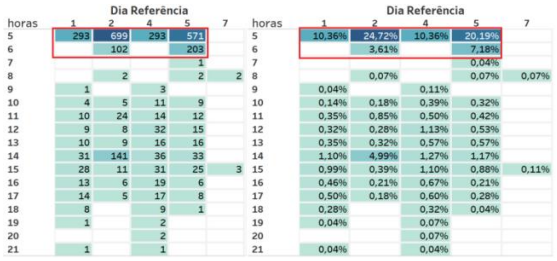
Relativamente a *Logging Advice*: 1) Número de chamadas aos serviços: Somar o número de chamadas aos serviços em cada hora; 2) Número de chamadas aos métodos: Somar o número de chamadas aos métodos em cada hora.
Quanto a CCCException temos: Número de erros por tipo: Somar os erros registados em cada hora por tipo de erro.

VIII. RESULTADOS E DISCUSSÃO

Os dados utilizados neste trabalho correspondem aos *logs* de 7 dias seguidos (uma semana completa) entre os dias 30-12-2019 e 5-1-2020: incluem um dia feriado e um fim de semana. As seguintes tabelas e representações gráficas resultaram da

análise efetuada aos dados recolhidos, tendo sido utilizadas as funcionalidades da ferramenta Tableau na geração das tabelas e gráficos de suporte à visualização de dados e o Excel para algumas tabelas e cálculos adicionais.

Session Manager: nas seguintes tabelas é possível visualizar o padrão horário das sessões perdidas no período em análise. Verifica-se que, aproximadamente, 75% das sessões perdidas ocorrem no período entre as 5 e as 7 horas com a exceção dos dias 3 e 6 em que não foram registas perdas. Uma análise detalhada aos *logs* permitiu verificar que a origem desta concentração de erros estava associada à paragem e arranque da instância do J-Boss que ocorre diariamente por volta da 5:10, provocando o fim abrupto de todas as sessões em curso e de todas as que sejam iniciadas durante o período de inatividade.

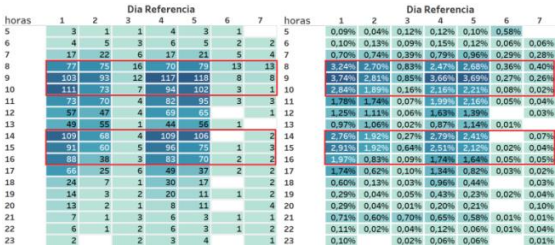


horas	Dia Referência 1	2	4	5	7
5	293	699	293	571	
6		102		203	
7				1	
8		2		2	2
9	1		3		
10	4	5	11	9	
11	10	24	14	12	
12	9	8	32	15	
13	10	9	16	16	
14	31	141	36	33	
15	28	11	31	25	3
16	13	6	19	6	
17	14	5	17	8	
18	8		9	1	
19	1		2		
20			2		
21	1		1		

horas	Dia Referência 1	2	4	5	7
5	10,36%	24,72%	10,36%	20,19%	
6		3,61%		7,18%	
7				0,04%	
8		0,07%		0,07%	0,07%
9	0,04%		0,11%		
10	0,14%	0,18%	0,39%	0,32%	
11	0,35%	0,85%	0,50%	0,42%	
12	0,32%	0,28%	1,13%	0,53%	
13	0,35%	0,32%	0,57%	0,57%	
14	1,10%	4,99%	1,27%	1,17%	
15	0,99%	0,39%	1,10%	0,88%	0,11%
16	0,46%	0,21%	0,67%	0,21%	
17	0,50%	0,18%	0,60%	0,28%	
18	0,28%		0,32%	0,04%	
19	0,04%		0,07%		
20			0,07%		
21	0,04%		0,04%		

Figura 8 - Sessões perdidas

As tabelas seguintes representam o padrão horário da criação de sessões (em número ou percentagem do total da semana) e verifica-se que aproximadamente 60% das sessões são iniciadas nos inícios dos períodos de trabalho (8 as 10 e 14 as 16).



horas	Dia Referencia 1	2	3	4	5	6	7
5	3	1	1	4	3	1	
6	4	5	3	6	5	2	2
7	17	22	6	17	21	5	4
8	77	75	16	70	73	13	13
9	103	93	12	117	118	8	8
10	111	73	7	94	102	3	1
11	73	70	4	82	95	3	3
12	57	47	4	69	65	1	
13	49	55	1	44	56	1	
14	109	68	4	109	106		2
15	91	60	5	96	75	1	3
16	88	38	3	83	70	2	2
17	66	25	6	49	37	2	2
18	24	7	1	30	17		2
19	14	3	2	20	11	1	2
20	13	2	1	8	11		4
21	7	1	3	6	3	1	1
22	6	1	2	6	3	1	2
23	2		2	3	4		1

horas	Dia Referencia 1	2	3	4	5	6	7
5	0,09%	0,04%	0,12%	0,12%	0,10%	0,58%	
6	0,10%	0,13%	0,09%	0,15%	0,12%	0,06%	0,06%
7	0,70%	0,74%	0,39%	0,79%	0,96%	0,29%	0,28%
8	3,26%	2,70%	0,83%	2,42%	2,68%	0,36%	0,40%
9	3,74%	2,81%	0,85%	3,66%	3,69%	0,27%	0,26%
10	2,84%	1,89%	0,16%	2,16%	2,21%	0,08%	0,02%
11	1,78%	1,74%	0,07%	1,99%	2,16%	0,05%	0,04%
12	1,25%	1,11%	0,06%	1,63%	1,39%		0,03%
13	0,97%	1,06%	0,02%	0,87%	1,14%	0,01%	
14	2,76%	1,92%	0,27%	2,72%	2,41%		0,07%
15	2,91%	1,92%	0,64%	2,51%	2,12%	0,02%	0,04%
16	1,97%	0,83%	0,09%	1,74%	1,64%	0,05%	0,05%
17	1,74%	0,62%	0,10%	1,34%	0,82%	0,03%	0,02%
18	0,60%	0,13%	0,03%	0,96%	0,44%		0,03%
19	0,29%	0,04%	0,05%	0,43%	0,23%	0,02%	0,04%
20	0,29%	0,04%	0,01%	0,20%	0,21%		0,10%
21	0,71%	0,60%	0,70%	0,65%	0,58%	0,01%	0,01%
22	0,11%	0,02%	0,04%	0,12%	0,06%	0,01%	0,04%
23	0,10%		0,02%	0,06%	0,06%		0,01%

Figura 9 - Sessões criadas

Este padrão é mais visível nos dias 1, 2, 4 e 5 já que os dias 3,6 e 7 correspondem, nesta semana em análise, respetivamente a um feriado, um sábado e um domingo. A utilização atípica do SI nos dias 3,6 e 7 fica clara no gráfico seguinte onde está representado o número de utilizadores que acedeu ao sistema em cada um dos dias do período em análise.

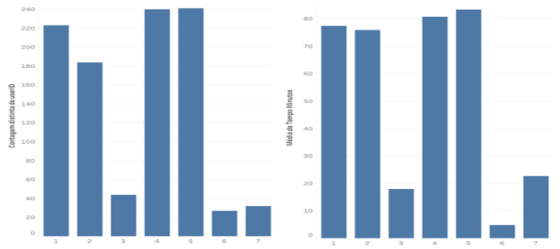


Figura 10 - Utilizadores por dia e Tempo médio por sessão

Em termos de utilização do SI, e considerando apenas os dias normais de trabalho, os gráficos apresentam pequenas variações quer seja em termos de utilizadores (221 utilizadores em média) quer seja em termos de tempos médios de utilização por sessão (80 minutos de duração em média) (Figura 10).

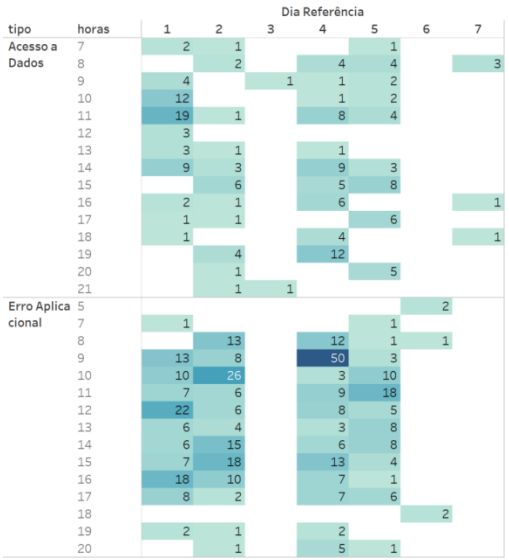
	Dia Referencia 1	2	3	4	5	6	7	Totais
Terminadas Normalmente	2629	1840	454	2473	2311	184	150	10041
Terminadas Abrutamente	423	1012		486	902		5	2828
	13,9%	35,5%	0,0%	16,4%	28,1%	0,0%	3,2%	22,0%

Figura 11 - Sessões perdidas versus Sessões Criadas

Comparando o número de sessões perdidas, face ao número total de sessões criadas (Terminadas Normalmente + Terminadas Abruptamente), resulta uma percentagem 22% (Figura 11), um valor significativo que é explicado pelo efeito da paragem e arranque da instância do J-Boss. Por ocorrer fora de horas e pelo facto de os processos automáticos correrem em ciclo até que terminem com sucesso a atividade, acaba por não ter impacto significativo na atividade dos utilizadores.

CCCException: na figura 12 é possível comparar os erros aplicacionais e os erros de acesso a dados, considerando o total de erros. Deste gráfico resulta uma clara prevalência dos erros aplicacionais face aos erros de acesso aos dados.

Figura 12 - Rácio entre tipos de erros



tipo	horas	Dia Referência 1	2	3	4	5	6	7
Acesso a Dados	7	2	1			1		
	8		2		4	4		3
	9	4		1	1	2		
	10	12			1	2		
	11	19	1		8	4		
	12	3						
	13	3	1		1			
	14	9	3		9	3		
	15		6		5	8		
	16	2	1		6			1
	17	1	1			6		
	18	1			4			1
	19		4		12			
	20		1			5		
	21		1	1				
Erro Aplicacional	5						2	
	7	1				1		
	8		13		12	1	1	
	9	13	8		50	3		
	10	10	26		3	10		
	11	7	6		9	18		
	12	22	6		8	5		
	13	6	4		3	8		
	14	6	15		6	8		
	15	7	18		13	4		
	16	18	10		7	1		
	17	8	2		7	6		
	18						2	
	19	2	1		2			
	20		1		5	1		

Figura 13 - Padrão Horário dos erros

Da análise dos padrões horários verifica-se que os erros se concentram nos dias úteis (1, 2, 4, 5) no período normal de trabalho (das 8 às 18) (Figura 13) o que está em linha com os períodos de maior utilização do SI (Figura 9).

Com base no gráfico da Figura 14 e na tabela da Figura 15 destacam-se claramente 5 serviços (Order/PrimaryLoad/ CreditDebit/Claim/VesselVoyage) que, em conjunto, representam aproximadamente, 75% dos erros registados. Estes serviços coincidem, como se vai verificar na análise das chamadas aos serviços, com alguns dos serviços mais utilizados.

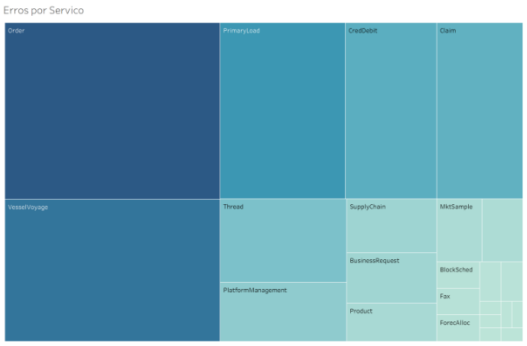


Figura 14 - Distribuição dos erros por serviços

	Dia Referência						
Service	1	2	3	4	5	6	7
Order	6,066%	9,185%	0,173%	4,333%	3,293%		
VesselVoyage	2,253%	4,506%		9,532%	2,253%		
PrimaryLoad	4,506%	1,213%		2,946%	4,333%		0,347%
CredDebit	4,679%	1,386%		3,120%	0,520%		
Claim	4,679%	2,080%		1,213%	1,213%		
Thread	1,213%	0,173%	0,173%	2,600%	1,906%	0,347%	
PlatformManageme..	0,173%	0,867%		2,080%	1,386%		
SupplyChain		2,946%					
BusinessRequest				2,600%	0,173%		
Product	1,560%	0,173%			0,347%		
MktSample					1,733%		
StoredProcedure	0,347%	0,173%		0,867%		0,173%	
BlockSched	0,520%				0,173%		
ForecAlloc				0,693%			
Fax	0,347%			0,347%			
Sales							0,520%
CCCMailSenderImpl		0,173%				0,347%	
TabLogistic	0,173%						
SGFTP	0,173%						
Rebate				0,173%			
Price	0,173%						
MarketingBudget					0,173%		
CreditControl	0,173%						

Figura 15 - Erros aplicacionais por serviço

	Dia Referência				
Service	1	2	4	5	6
VesselVoyage	3,20%	6,40%	13,55%	3,20%	
BusinessRequest			3,69%	0,25%	
Order	6,65%	11,82%	3,45%	4,19%	
Thread	0,99%		2,22%	2,22%	0,49%
CredDebit	4,93%	0,74%	1,97%	0,74%	
PlatformManageme..	0,25%	0,25%	1,72%	0,74%	
Claim	3,45%	2,46%	1,72%	1,48%	
ForecAlloc			0,99%		
PrimaryLoad	1,48%	0,74%	0,74%	0,25%	
Fax	0,49%		0,49%		
Rebate			0,25%		
SupplyChain		4,19%			
Product	2,22%			0,49%	
MktSample				2,46%	
StoredProcedure	0,49%	0,25%			0,25%
CCCMailSenderImpl		0,25%			0,49%
TabLogistic	0,25%				
SGFTP	0,25%				
MarketingBudget				0,25%	

Figura 16 - Erros aplicacionais por serviço

	Dia Referência					
Service	1	2	3	4	5	7
PrimaryLoad	11,70%	2,34%		8,19%	14,04%	1,17%
Claim	7,60%	1,17%			0,58%	
Order	4,68%	2,92%	0,58%	6,43%	1,17%	
CredDebit	4,09%	2,92%		5,85%		
BlockSched	1,75%				0,58%	
Thread	1,75%	0,58%	0,58%	3,51%	1,17%	
CreditControl	0,58%					
Price	0,58%					
PlatformManageme..		2,34%		2,92%	2,92%	
Product		0,58%				
StoredProcedure				2,92%		
Sales						1,75%

Figura 17 - Erros de acesso a dados por serviço

	Dia Referencia							
	1	2	3	4	5	6	7	Totais
Erros	253	159	2	176	104	5	5	704
Chamadas	198 753	136 770	6 702	182 202	145 944	2 807	2 647	675 825
%	0,127%	0,116%	0,030%	0,097%	0,071%	0,178%	0,189%	0,104%

Figura 18 - Erros registados versus Chamadas aos serviços

Comparando os erros aplicacionais com os erros de acesso a dados verificam-se diferenças relevantes ao nível dos serviços que são afetados por cada um dos tipos. Comparando o número de erros registados com o total das chamadas efetuadas, no mesmo período, resulta uma percentagem de 0,104 %, o que indicia uma boa fiabilidade e disponibilidade do SI (Figura 18). Logging Advice: na figura 19 é possível verificar que os métodos associados a Pesquisa de Dados são sempre superiores.

	Dia Referência						
Tipo Servico	1	2	3	4	5	6	7
Pesquisa Dados	84,21%	84,79%	92,64%	84,07%	85,10%	75,63%	84,74%
Actualiza Dados	7,29%	6,21%	6,34%	6,47%	6,25%	2,35%	13,56%
Valida Dados	3,23%	3,18%		3,66%	3,19%	9,16%	0,34%
Calcula Dados	2,85%	3,08%	0,04%	3,11%	2,81%	6,52%	0,30%
Nao Classificado	1,38%	1,48%	0,31%	1,69%	1,53%	5,77%	0,04%
Gerar Documentos	0,81%	1,06%	0,37%	0,75%	0,81%		
Menu	0,24%	0,21%	0,28%	0,25%	0,31%	0,57%	1,02%

	Dia Referência						
Tipo Servico	1	2	3	4	5	6	7
Pesquisa Dados	167 360	115 965	6 209	153 179	124 194	2 123	2 243
Actualiza Dados	14 482	8 487	425	11 788	9 122	66	359
Valida Dados	6 421	4 346		6 666	4 660	257	9
Calcula Dados	5 668	4 212	3	5 674	4 102	183	8
Nao Classificado	2 737	2 019	21	3 076	2 230	162	1
Gerar Documentos	1 614	1 454	25	1 372	1 184		
Menu	471	287	19	447	452	16	27

Figura 19 - Chamadas por tipo de método

Identificam-se os 5 serviços que, em termos de utilização, se destacam dos demais. Comparando esta informação com a dos erros por serviço, com exceção do VesselVoyage, existe uma associação entre os erros e a respetiva utilização.

Service	1	2	3	4	5	6	7
Order	48 549	33 420	226	44 096	34 864	1 229	137
PrimaryLoad	36 764	24 870	1 756	26 552	26 462	23	1 487
ShipDocs	25 016	17 358		24 561	22 620		19
StoredProcedure	20 562	14 224	531	19 350	15 263	264	379
CredDebit	7 894	6 716	2 573	7 775	4 796	53	14
Claim	7 760	3 500		3 080	1 699		
Lookup	5 158	3 355	2	5 201	3 454	120	21
Product	4 059	2 738	3	3 784	3 126	90	14
BusinessUnit	4 048	2 623	100	3 565	3 018	56	72
Destination	2 949	1 976	1	2 960	2 159	115	

Service	1	2	3	4	5	6	7
Order	7,184%	4,945%	0,033%	6,525%	5,159%	0,182%	0,020%
PrimaryLoad	5,440%	3,680%	0,260%	3,929%	3,916%	0,003%	0,220%
ShipDocs	3,702%	2,568%		3,634%	3,347%		0,003%
StoredProcedure	3,043%	2,105%	0,079%	2,863%	2,258%	0,039%	0,056%
CredDebit	1,168%	0,994%	0,381%	1,150%	0,710%	0,008%	0,002%
Claim	1,148%	0,518%		0,456%	0,251%		
Lookup	0,763%	0,496%	0,000%	0,770%	0,511%	0,018%	0,003%
Product	0,601%	0,405%	0,000%	0,560%	0,463%	0,013%	0,002%
BusinessUnit	0,599%	0,388%	0,015%	0,528%	0,447%	0,008%	0,011%
Destination	0,436%	0,292%	0,000%	0,438%	0,319%	0,017%	

Figura 20- Top 10 dos serviços utilizados

Criou-se um *Dashboard* resumo da atividade mais relevante no período de análise. A Figura 21 apresenta a informação mais relevante por dia: 1) N.º total de utilizadores registados; 2) N.º total de chamadas aos serviços efetuadas; 3) Tempo médio por sessão em minutos; 4) Total de erros registados. A Figura 22 apresenta os padrões horários semanais: 1) N.º total de utilizadores registados em cada hora; 2) N.º total de chamadas aos serviços em cada hora; 3) N.º total de sessões perdidas em cada hora; 4) N.º total de erros registados em cada hora. A Figura 23 apresenta o resumo dos utilizadores e sessões: 1) N.º total de utilizadores registados em cada hora; 2) N.º total de utilizadores registados no dia; 3) Tempo médio por sessão em minutos. A Figura 24 destaca o resumo dos erros registados: 1) Percentagem dos erros registados na hora face ao total da semana; 2) Percentagem dos erros registados serviço a serviço face ao total da semana; 3) Análise comparativa entre % semanais dos erros aplicacionais face aos erros de acesso a dados. Na Figura 25 consta a informação resumo do número total de acessos a dados por hora, dia a dia, das chamadas efetuadas por cada um dos seguintes tipos: 1) Acesso a dados; 2) Atualização de dados:

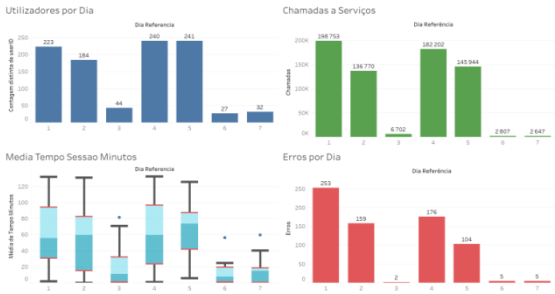


Figura 21 - Dashboard Resumo

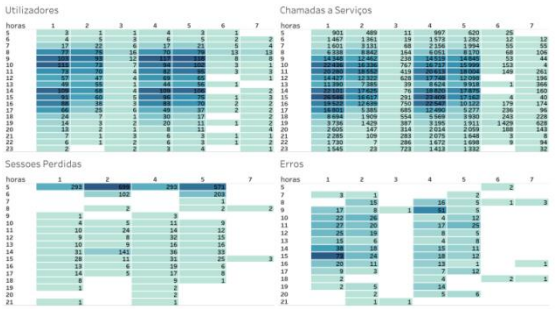


Figura 22 - Padrões horários de utilização

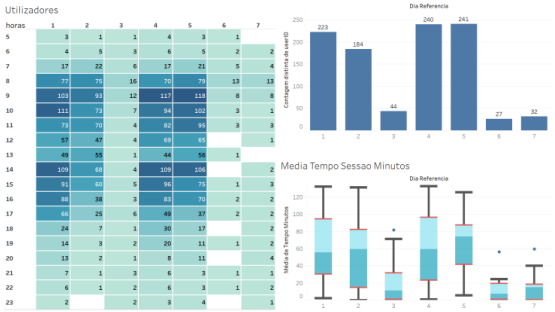


Figura 23- Utilizadores / Acessos

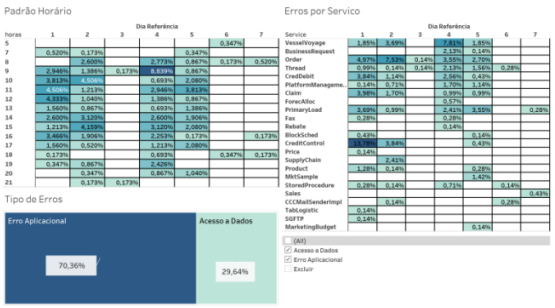


Figura 24 – Erros

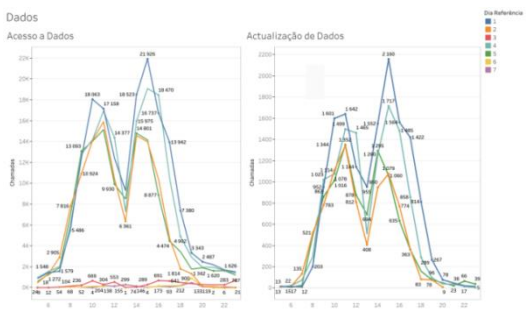


Figura 25 - Dados

IX. CONCLUSÕES

O processo de extração de dados correspondeu à parte mais complexa e demorada do trabalho por ter obrigado à criação de um programa de raiz (aproximadamente 15 horas). É importante ressalvar que este esforço, em futuras análises, não será necessário, mas apenas considerar o tempo de processamento dos respetivos ficheiros. Os dados obtidos permitem uma visão global e bastante detalhada da utilização do SI. Foi particularmente interessante verificar os padrões de utilização, quer seja em termos temporais (dias e horas) quer em termos funcionais. Esta informação será útil no planeamento das intervenções ao SI para reduzir o impacto nas atividades dos utilizadores. Outra conclusão interessante está relacionada com o rácio de 10/1 observado entre os serviços de Acesso a Dados e os restantes tipos o que permite otimizar a configuração da BD de forma a melhorar a sua performance.

A impossibilidade de, com base na informação dos *logs*, estabelecer relação entre as chamadas aos serviços e as respetivas sessões, bem como de estabelecer uma relação entre os erros e as respetivas sessões, não permitiu uma análise conjunta dessas informações. A inexistência de registo das respostas aos serviços não permitiu fazer uma análise da capacidade de resposta do servidor SOA. Analisar a evolução da resposta do servidor SOA ao longo do dia e a análise do efeito do número de chamadas efetuadas aos serviços na sua capacidade de resposta são objetivos que devem ser tidos em consideração no futuro caso seja possível enriquecer o conteúdo dos dados existentes nos ficheiros de *logs*. Para colmatar estas falhas deve resultar deste trabalho a redefinição do registo de *logs* para que: todos os registos tenham associado o respetivo session ID; incluir nos *logs* o registo das respostas aos serviços. Uma vez implementadas estas alterações, será necessário rever: 1) o processo ETL; 2) a estrutura de dados para receber a nova informação; 3) tratamento estatístico para definir novas métricas de controlo e novos quadros de análise.

Como base nestes resultados confirma-se a possibilidade de extrair informação relevante a partir de ficheiros de *logs* aplicacionais, sendo que essa pode ser utilizada em tratamento estatístico, permitindo uma visão, global e detalhada, da forma como o sistema é utilizado. A ausência de métricas de referência não permitiu efetuar uma análise comparativa com outros trabalhos, o que propomos como trabalho futuro.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Hakan Yilmaz AND Pinar Senkul "Using Ontology and Sequence Information for Extracting Behavior Patterns from Web Navigation Logs",2010 IEEE International Conference on Data Mining Workshops, 2010.

[2] Aditi Shrivastava, Nitin Shukla "Extracting Knowledge from User Access Logs", International Journal of Scientific and Research Publications, April 2012

[3] Jianyi Wang,Lihong Jiang AND Hongming Cai "Scenario-based Method for Business Process Analysis and Improvement in SOA", 2014 IEEE 11th International Conference on e-Business Engineering, 2014

[4] Borges, L.C., Marques, V.M., Bernardino, J.,Comparison of data mining techniques and tools for data classification, (2013) ACM International Conference Proceeding Series, pp. 113-116