Integração e Análise das Despesas do Governo Federal

Cristian Weiland[1], **Diego Pasqualin**[1],
Edemir Maciel[1], **Luis C. E. de Bona**[1], **Marcos Sunye**[1]

[1]Centro de Computação Científica e Software Livre
Departamento de Informática
Universidade Federal do Paraná (UFPR)
R. Cel. Francisco H. dos Santos, 100 – Curitiba – PR – Brasil

{cw14,dpasqualin,bona,sunye,edemir}@inf.ufpr.br

Abstract. *Since 2012, with the creation of the law on access of information in Brazil, the amount of data available grew in such a way that the difficulty shifted from the access itself to the treatment and interpretation of the available information. In this context, SIMTransparência was developed with the objective of providing an alternative view of the Portal of Transparency of the Federal Government, displaying data from several datasets in an intuitive and dynamic interface, capable of assisting the public manager and brazilian citizens to better understand and control expenditures of the federal government. This article describes the technologies within SIMTransparência and ways to replicate it's datasets.*

Resumo. *Desde 2012, com o início da vigência da Lei de Acesso a Informação, a quantidade de dados disponíveis cresceu de tal forma que a dificuldade passou a ser não o acesso em si, mas o tratamento e interpretação das informações disponibilizadas. Nesse contexto o SIMTransparência foi desenvolvido, com o objetivo de proporcionar uma visualização alternativa ao Portal da Transparência do Governo Federal, exibindo dados de várias bases em uma interface intuitiva e dinâmica, capaz de auxiliar o gestor e cidadão comum a melhor compreender e fiscalizar as despesas do governo federal. Esse artigo descreve as tecnologias envolvidas no desenvolvimento do SIMTransparência e formas de replicar sua base de dados.*

1. Introdução

A Lei de Acesso a Informação (12.527/2011) garante a qualquer cidadão o acesso a informações sobre as atividades de órgãos públicos, sem a necessidade de identificação ou motivo. Desde o início de vigência da lei, em 2012, a quantidade de informação disponibilizada vem crescendo consideravelmente, reduzindo o problema de acesso aos dados, mas intensificando o problema de análise, a qual podemos considerar inviável sem o auxílio de ferramentas computacionais capazes de filtrar, manipular e destacar dados de interesse.

O SIMTransparência[1] é um projeto de Software Livre criado com esse propósito. Através da manipulação de uma grande quantidade de dados e exibição intuitiva em in-

[1]Todo o código fonte, assim como documentação detalhada, encontra-se disponível sob licença GPL no repositório `https://gitlab.c3sl.ufpr.br/c3sl/transparencia`.

terface gráfica disponível via internet, ele busca aproximar cidadãos e gestores do processo de fiscalização dos gastos públicos. Atualmente, todos os dados disponibilizados pelo SIMTransparência são periodicamente coletados, tratados e filtrados a partir do Portal da Transparência do Governo Federal[2]. Ao grupar várias bases de dados e permitir aplicação de filtros avançados em diferentes escalas de tempo, o SIMTransparência se apresenta como uma ferramenta dinâmica, complementar ao Portal da Transparência e capaz de evidenciar despesas controversas e prioridades do órgãos do governo federal nos gastos públicos. Os dados coletados são tratados por *scripts* em *Shell* e *Python*, armazenados no banco de dados ElasticSearch[3] e, por fim, visualizados em um navegador com a utilização da ferramenta Kibana[3]. Uma instância do projeto está disponível publicamente no endereço `http://www.c3sl.ufpr.br/transparencia`. Ela pode ser facilmente replicada e modificada em outros ambientes, processo este objetivo primário desse artigo.

As próximas seções descrevem o processo necessário para replicar o SIMTransparência em um ambiente GNU/Linux. A seção 2 detalha as bases de dados que serão utilizadas. As características do banco de dados são descritas na seção 3, enquanto que a seção 4 descreve o método sugerido para visualização dos dados armazenados. Em seguida, na seção 5, um resumo é apresentado, com os passos necessários para colocar o ambiente em produção, seguido por fim os desafios e limitações do sistema (seção 7) e a conclusão do trabalho.

2. Bases de Dados

Ao todo, sete conjuntos de dados do Portal da Transparência são utilizados: **i)** Diárias de Viagens[4], concedidas ao servidor civil por dia de afastamento da sede do serviço, destinando-se a indenizá-lo por despesas extraordinárias com pousada, alimentação e locomoção urbana. **ii)** Cadastro de Servidores[5], dados cadastrais de cada servidor, incluindo identificação por nome e CPF mascarado, cargo, função e atividades, entidade de lotação e de exercício, além do tipo e situação do vínculo, informações sobre afastamento (se houver), datas e documentos referentes ao ingresso e à jornada de trabalho. **iii)** Remuneração de Servidores[5], corresponde remuneração e subsídio recebidos por ocupante de cargo e emprego público, incluindo auxílios, ajudas de custo, jetons e quaisquer outras vantagens pecuniárias, bem como proventos de aposentadoria e pensões. **iv)** Gastos Diretos[6], despesas de custeio de manutenção das atividades dos órgãos da administração pública, como por exemplo: despesas com pessoal, juros da dívida, aquisição de bens de consumo, serviços de terceiros, manutenção de equipamentos, obras, água, energia, telefone, dentre outros. **v)** Lista de favorecidos por Natureza Jurídica[7], lista contendo possíveis regimes jurídicos atribuídos a uma empresa ou agente autônomos (ex: autar-

[2]Site do Portal da Transparência: `http://portaltransparencia.gov.br/`.

[3]Site oficial da Elastic, empresa responsável pelo ElasticStack: `https://www.elastic.co/`.

[4]Diárias de viagens: `http://www.portaltransparencia.gov.br/downloads/mensal.asp?c=Diarias`.

[5]Cadastro de servidores e remuneração: `http://www.portaltransparencia.gov.br/downloads/servidores.asp`.

[6]Gastos diretos: `http://www.portaltransparencia.gov.br/downloads/mensal.asp?c=GastosDiretos`.

[7]Natureza jurídica, favorecidos e CNAE: `http://www.portaltransparencia.gov.br/downloads/mensal.asp?c=FavorecidosTransferencias`.

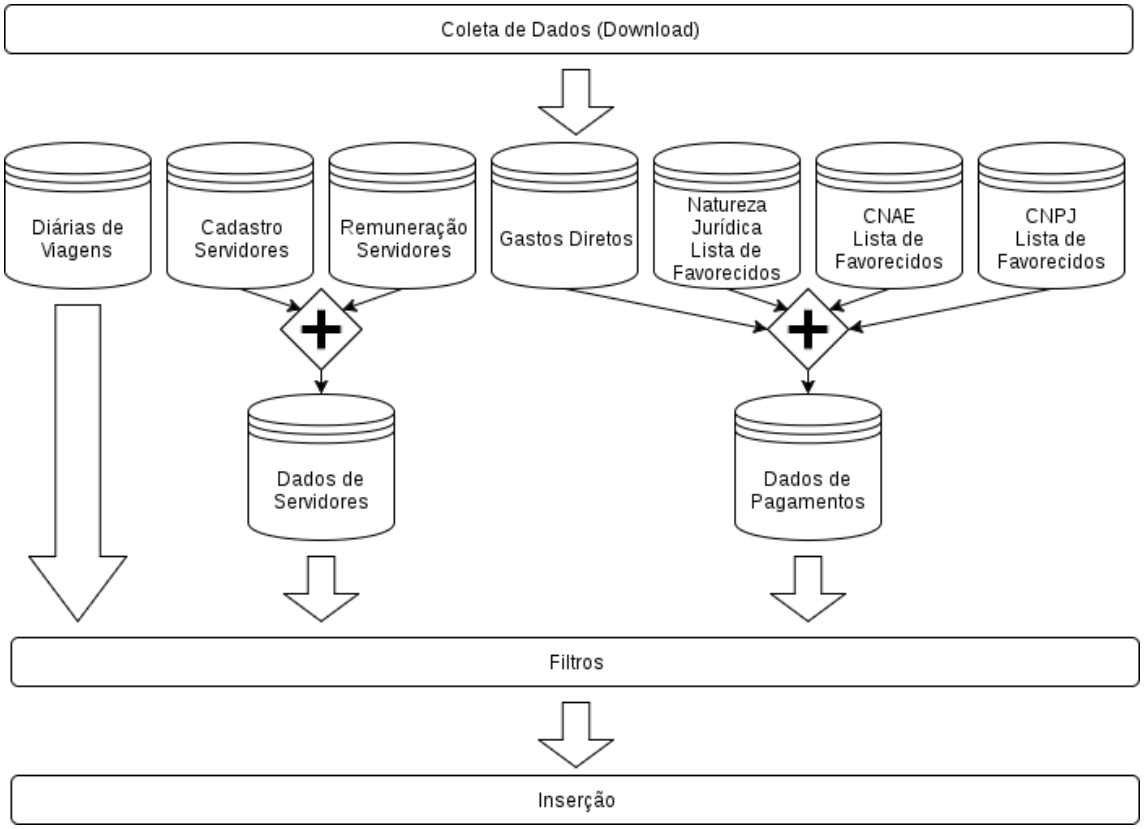

Figura 1. Visão geral do processo de coleta, tratamento e armazenamento dos dados do SIMTransparência.

quia, sociedade anônima, empresa individual, etc). **vi)** Lista de favorecidos por CNAE[7], contendo nome de todas as atividades econômicas extraídas do cadastro da Classificação Nacional de Atividades Econômicas[8]. **vii)** Lista de favorecidos por CNPJ[7], lista dos favorecidos - pessoa jurídica - dos pagamentos realizados pelos órgãos e entidades da Administração Pública Federal, que executam as despesas pelo Sistema Integrado de Administração Financeira do Governo Federal (Siafi).

Conforme exibido na Figura 1, as bases de "Natureza Jurídica", "CNAE" e "CNPJ" são agrupadas em "Dados de Pagamentos", com o objetivo de adicionar no conjunto de "Gastos Diretos" o setor de atuação e natureza jurídica das empresas favorecidas, informação que facilita a identificação das áreas de atuação (energia, administração, limpeza, etc) com gastos mais expressivos. A união entre "Cadastro Servidores" e "Remuneração Servidores", por sua vez, permite visualizar dados pessoais dos servidores públicos (como nome, profissão e alocação) junto com sua respectiva remuneração. É importante salientar que a tabela de "Cadastro de Servidores" possui uma entrada para cada cargo ocupado pelo servidor (ex: professor e chefe de departamento) e, nesses casos, os dados de remuneração, que são disponibilizados como um único valor independente dos cargos ocupados pelo servidor, são preenchidos em apenas uma das entradas, com as demais recebendo valores nulos. Esse procedimento elimina problemas de dupla contagem em agregações que envolvam a remuneração de servidores.

[8]CNAE: `http://concla.ibge.gov.br/`.

A coleta é feita através de requisições *http* para o Portal Transparência do Governo Federal, que possui *urls* padronizadas de acordo com o conjunto de dados e data de disponibilização, com periodicidade mensal para a maior parte dos conjuntos. Como resultado, obtém-se arquivos *CSV*, que podem ser filtrados e armazenados de acordo com necessidade do usuário e capacidade computacional disponível.

3. Armazenamento

O armazenamento é feito no ElasticSearch, um banco de dados documental, de código aberto, distribuído e escalável, com capacidade de gerir grandes volumes de dados e fazer agregações de forma eficiente [Kononenko et al. 2014]. Ele tem como unidade básica um documento no formato JSON e documentos podem ser agrupados em conjuntos, formando índices.

O ElasticSearch permite a fragmentação dos dados dos índices em *shards*, que são distribuídos entre as máquinas do *cluster*, aumentando assim a eficiência das buscas e agregações. Por padrão, cada índice é separado em cinco *shards*, cada qual com uma réplica, totalizando dez fragmentos por índice. É importante considerar que a criação excessiva de índices implica em um alto custo computacional no gerenciamento dos fragmentos, ao passo que um número reduzido subutiliza o potencial de processamento paralelo e distribuído do ElasticSearch. No SIMTransparência, os índices são particionados pela fonte de dados (despesas, servidores e diárias), ministério (ex: MEC, MCTIC), além do mês e o ano em que os dados foram disponibilizados, gerando índices como `despesas-mec-2017-12` e `despesas-mctic-2014-03`. Acreditamos que esse modelo representa um crescimento controlado e com performance satisfatória.

4. Visualização

A visualização dos dados é feita com o Kibana, ferramenta desenvolvida para uso conjunto com o ElasticSearch. No Kibana, o administrador pode criar uma grande variedade de visualizações, desde gráficos de barra e pizza, até tabelas, mapas geográficos e de calor. Visualizações relacionadas compõe um *dashboard*, que pode ser compartilhado em páginas externas. O bom aproveitamento da ferramenta se dá através da criação de *dashboards* compreensivos formados por visualizações complementares, que permitam a exploração do mecanismo de filtros disponibilizado pela ferramenta.

Os filtros são aplicáveis em qualquer visualização através da seleção de um elemento de interesse, ação que propaga este mesmo filtro em todas as demais visualizações do *dashboard*. Tal mecanismo permite um número elevado de combinações que destaca, a cada interação, uma nova característica do conjunto de dados em análise. Por exemplo, no *dashboard* “MEC - Despesas” do SIMTransparência, podemos filtrar no gráfico “Descrição Seção” o elemento “Limpeza em Prédios e em Domicílios” e, em seguida, selecionar algum favorecido no gráfico ao lado, que já apresenta somente empresas relacionadas a limpeza. Mais acima, a tabela “Lista de Universidades” informa então quem possui o maior contrato com o favorecido selecionado. O usuário pode continuar a pesquisa habilitando novos filtros relacionados a data, elemento de despesa, ou outros, destacando situações de interesse em análise dinâmica. O repositório do SIMTransparência disponibiliza um arquivo com a definição dos elementos visuais e *dashboards* para serem carregados no Kibana (seção 5).

5. Instalação e Carregamento

Para replicar o SIMTransparência os seguintes passos são necessários: **i)** baixar o repositório do SIMTransparência, **ii)** opcionalmente instalar o ElasticStack (Logstash, Kibana e Elasticsearch), **iii)** executar programa que busca e carrega as bases de dados, **iv)** criar índices e **iv)** importar as visualizações e *dashboards* no Kibana.

O código fonte do SIMTransparência encontra-se no repositório GIT `https://gitlab.c3sl.ufpr.br/c3sl/transparencia.git`. Após baixar o repositório, siga as instruções presentes no arquivo `README.md` para criar exemplo contendo as bases de dados de despesas, diárias e servidores do Ministério da Educação. Em seguida, opcionalmente instale a suíte do ElasticStack da forma descrita na documentação e execute o comando abaixo[9]:

```
# ./simtransparencia --todas-as-bases --inicio 2017-01 --fim 2017-12
```

Após término da execução, crie os índices na interface gráfica do Kibana acessando-o pelo navegador (por padrão, em `http://localhost:5601`) e clicando no botão "Management" no menu esquerdo, em seguida em "Index Patterns" e "Create Index Patterns". Insira então o valor `despesas-pagamentos-mec-*` e em "Time-field name" selecione `Data Pagamento Timestamp`. Repita o processo para criar os índices `despesas-diarias-mec-*` e `servidores-mec-*`. Por fim, para criar os gráficos e *dashboards* no Kibana, clique novamente em "Management", "Saved Objetcts", depois em "Import" no canto superior direito, selecionando por fim o arquivo `kibana_dashboards.json`, presente na raiz do repositório. Os *dashboards* podem agora ser acessados através do menu lateral esquerdo. Se preferível, é possível adicionar o parâmetro `--sem-elastic` e dispensar a ElasticStack. O programa irá gerar um arquivo CSV com os dados integrados que pode ser analisado com qualquer outra ferramenta.

6. Exemplos

Nesta seção será apresentado exemplo de visualização. Os dados são adquiridos sem modificação de valores a partir do Portal da Transparência do Governo Federal e a sua interpretação é de responsabilidade do leitor.

Um exemplo do gráfico de "Maiores favorecidos" pode ser visto na Figura 2. É notável que entre os dez maiores favorecidos, três são bancos. As instituições financeiras agem como intermediários em certas transações e, dessa forma, mascaram o destino final dos recursos. A mesma figura destaca também a participação da empresa "Sulclean" nas despesas da "Universidade Federal de Santa Maria" que, somadas as divisões de serviços gerais e segurança, corresponde a 25.61% do orçamento total da Universidade.

7. Desafios e limitações

Um dos desafios do SIMTransparência é disponibilizar os dados de toda a união, mas o volume de dados disponível requer uma grande quantidade de recursos computacionais. Essa limitação é atualmente mitigada pelo carregamento individual de apenas alguns ministérios, além da separação dos gráficos e tabelas em diferentes *dashboards*, o que reduz o número de consultas executadas no banco de dados a cada interação com o usuário. Nesse sentido, testes de performance e estabilidade são trabalhos futuros necessários para permitir o uso da ferramenta em larga escala através da interface publicamente acessível.

A interface do Kibana, apesar de proporcionar um grande avanço na visualização e navegabilidade se comparada ao atual Portal da Transparência, ainda possui diversas limitações.

[9] `# ./simtransparencia --help` exibe ajuda sobre os argumentos de linha de comando.

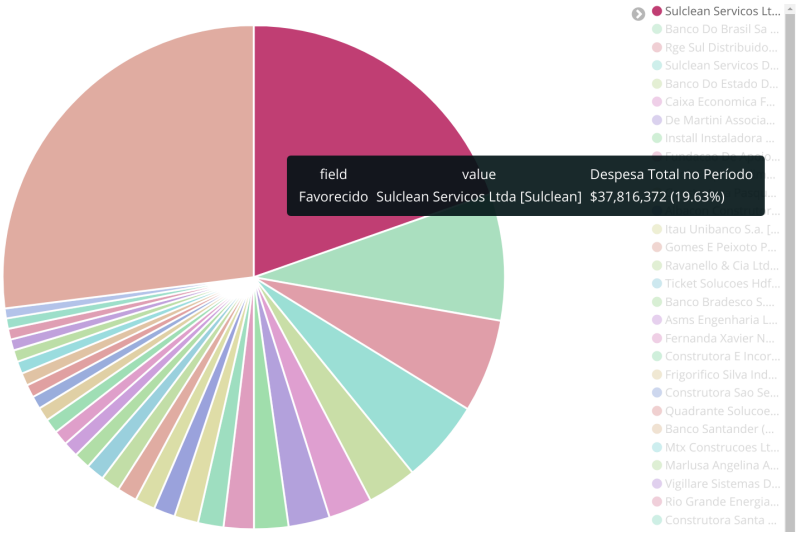

Figura 2. Gráfico exibindo maiores favorecidos com despesas na "Universidade Federal de Santa Maria".

Ela não é internacionalizável, os filtros aplicados não estão sempre visíveis e, mais importante, operações de interjeição são limitadas a apenas alguns tipos de visualizações. O Elasticsearch, por sua vez, não permite junção entre diferentes índices e em determinadas situações[10] calcula somente o valor aproximado das consultas, o que pode precisa ser tratado com cuidado em nosso caso de uso.

Por fim, os dados disponibilizados pelo Portal da Transparência possuem também limitações que, em parte, dificultam a análise. É possível observar, por exemplo, que bancos e fundações das Universidades são intermediários nas despesas da União, impossibilitando a identificação do destino final dos recursos aplicados. Ainda mais relevante, muitas entradas que são exibidas no Portal da Transparência não estão presentes nos arquivos CSV correspondentes disponibilizados para download, como o pagamento de mais de 458 bilhões em 2017 destinados a "Amortização e Juros da Dívida - Principal Corrigido da Dívida Mobiliária Refinanciado"[11].

8. Conclusão

Este artigo descreveu a combinação de uma série de bases de dados do Portal da Transparência do Governo Federal, com a geração de arquivos CSV e análise através da ferramenta gráfica Kibana. O SIMTransparência se coloca como uma visualização complementar aos dados disponibilizados no Portal da Transparência do Governo Federal, proporcionando algumas vantagens com relação a esse, como a liberdade para escolha do intervalo de tempo para análise e agrupamento de fontes diferentes em um único local. As diferentes visualizações disponibilizadas junto do repositório oficial permitem a gestores e usuários leigos a identificação rápida de despesas de maior relevância, proporcionando efetivamente a melhoria na transparência pública.

Referências

Kononenko, O., Baysal, O., Holmes, R., and Godfrey, M. W. (2014). Mining modern repositories with elasticsearch. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, pages 328–331. ACM.

[10]Elasticsearch e as consultas de valor aproximado: `https://tinyurl.com/mh4yhb7`.
[11]Link para o Portal da Transparência: `https://tinyurl.com/y7kcsavb`.