ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

# TRANSFORMANDO DADOS PÚBLICOS SEMIESTRUTURADOS EM UM BANCO DE DADOS DE DOCUMENTOS NOSQL

Tassio Sirqueira[1]
Jessica Faciroli[2]
Caroline Batista[3]
Pedro Coquito[4]

## RESUMO

Atualmente a Plataforma ECA busca auxiliar na extração de amostras de dados econômicos e em algumas análises voltadas para os economistas, focando em dados públicos dos programas sociais do Governo Federal. O Governo Federal disponibiliza mensalmente dados de todos os programas sociais em formato CSV (*Comma-separated values*), contudo, estudos econômicos normalmente envolvem a análise de dados complexos, fazendo uso de ferramentas estatísticas, onde amostras devem ser elaboradas. O objetivo deste trabalho é apresentar a ferramenta de transformação de dados de CSV para documentos JSON (JavaScript Object Notation), que alimenta a Plataforma ECA. O importado da Plataforma ECA é responsável por transformar todos os dados do Governo Federal de CSV para o

[1] Doutorando em Informática pela PUC-Rio, mestre em Ciência da Computação pela UFJF, bacharel em Sistemas de Informação pelo CES-JF e Professor dos cursos de ADS e SI nas Faculdades Integradas Vianna Júnior. E-mail: tmsirqueira@vianna.edu.br
[2] Doutoranda em Economia pela UFJF, mestre em Economia Aplicada pela UFJF, bacharel em Ciências Econômicas pela UFJF e Professora do Departamento de Economia da UFJF. E-mail: jessica.faciroli@economia.ufjf.br
[3] Aluna do curso de Sistemas para Internet pelas Faculdades Integradas Vianna Júnior. E-mail: caroline.batista@viannasempre.com.br
[4] Aluno do curso de Análise e Desenvolvimento de Sistemas pelas Faculdades Integradas Vianna Júnior. E-mail: pedro.coquito@viannasempre.com.br

ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

SGBD MongoDB, utilizando-se de uma aplicação de IO implementada em Java e Python.

**PALAVRAS-CHAVE: ECONOMIA COMPUTACIONAL. PROGRAMA BOLSA FAMÍLIA. DADOS PÚBLICOS. PROGRAMAS SOCIAIS. PLATAFORMA ECA.**

## INTRODUÇÃO

Mensalmente o Governo Federal do Brasil disponibiliza no portal da transparência[5] dados sobre os programas sociais, vinculados ao Cadastro Único (CadÚnico[6]), contendo diversas informações, permitindo assim que pesquisadores utilizem esses dados publicamente.

O Governo Federal possui atualmente 22 programas sociais que fazem uso da base de dados do CadÚnico, tendo como foco, as famílias que estão em situação de extrema pobreza e pobreza; sendo essa parcela da população atendida pelo programa Bolsa Família (PBF), conforme (BRASIL, 2018). Com esses dados disponibilizados, surgem diversas oportunidades de pesquisas principalmente para os economistas, porém para alcançar os resultados são necessários recursos computacionais, visto que esses dados são volumosos, criando o chamado Big Data.

Segundo Sosa Escurdero (2008), a revolução do Big Data está mudando radicalmente a forma como os dados são produzidos, gerenciados, armazenados e analisados pelos economistas, tornando mais acessíveis abordagens estatísticas e computacionais comumente utilizadas em outros campos, mas pouco exploradas na economia. Além disso, Sirqueira & Dalpra (2018) abordam que os desafios do Big

[5] Portal da Transparência - http://www.portaldatransparencia.gov.br/

[6] Cadastro Único - http://mds.gov.br/assuntos/cadastro-unico

ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

Data vão além, devido o volume de dados a ser analisado, a velocidade necessária para o processamento dos mesmos e em muitos casos a falta de estruturas fixas dos dados, dificultando a extração de informações úteis aos usuários.

A Plataforma *EconomiC Analyzer* (ECA) (SIRQUEIRA *et al*., 2018), busca auxiliar os pesquisadores da área de economia a manipular grandes volumes de dados semiestruturados, permitindo a extração de amostras, geração de relatórios em PDFs e gráficos. Atualmente a plataforma ECA encontra-se em teste pelo programa de pós-graduação em Economia da Universidade Federal de Juiz de Fora, auxiliando nas pesquisas que envolvem dados sobre o programa Bolsa Família.

Além desta introdução, na seção 2 abordaremos as características do Big Data e seu impacto na área de economia, desenvolvendo uma nova área denominada economia computacional. Na seção 3 apresentaremos detalhes da implementação do importador ECA, responsável por transformar os dados semiestruturados e não relacionais para documentos JSON que podem ser gerenciados pelo SGBD NoSQL MongoDB. Por fim, na seção 4 abordaremos alguns resultados obtidos, nossas considerações finais e os trabalhos futuros que serão desenvolvidos.

## 1 Big Data e Economia Computacional

Big Data corresponde não apenas a ideia de grandes volumes de dados, mas também a grandes variabilidades de formatos, necessidade de velocidade de processamento e a complexidade envolvida ao dado, requerendo tecnologias e técnicas avançadas em engenharia de dados e de software para capturar, armazenar e analisar as informações, conforme Sirqueira & Dalpra (2018).

No Big Data foram atribuídas características ao conceito consideradas os 5 V’s do Big Data: i) volume, ii) variedade, iii) velocidade, iv) veracidade e v) valor. O

ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

volume refere-se à quantidade de dados gerados como um todo, onde segundo McAfee *et al.* (2012), a cada segundo circulam na internet um número de dados superiores as informações armazenadas na rede há vinte anos atrás, o que permite empresas e o governo trabalharem com várias informações de diversas fontes.

Gandomi & Haider (2015) definem que a variedade consiste na heterogeneidade estrutural do conjunto de dados, ou seja, é a utilização de vários tipos de dados, englobando os estruturados, semiestruturados e não estruturados. Para os mesmos autores, a velocidade refere-se à velocidade na qual os dados são gerados e disponibilizados para análise.

A velocidade impacta na qualidade dos dados e informações geradas, com o objetivo de trazer informações autênticas que contribuam para uma correta tomada de decisão ao final de um estudo (LOBO, 2017).

Por fim, o valor pode ser dividido em valor do custo da tecnologia e valor derivado do uso do Big Data e valor do dado. O valor do custo da tecnologia é o custo necessário para obter uma certa tecnologia de hardware ou software, e o custo do Big Data refere-se a custo de capital, como custo da eficiência operacional e melhorias de procedimentos de negócios, e o valor do dado o quanto aquele dado é importante para o estudo e o que ele representa.

Para trabalhar com o Big Data, na ciência econômica está surgindo o campo denominado economia computacional, é definida pela *Society for Computational Economics*[7] (SCE), como a exploração da interseção da economia com a computação.

Visando auxiliar o pesquisador, a plataforma *EconomiC Analyzer* possibilita a coleta e seleção dos dados, além da geração de análises básicas, dispensando o uso de ferramentas atualmente utilizadas pelos economistas como SPSS, R e STATA. Os dados disponibilizados no portal da transparência pelo Governo Federal seguem o formato CVS, separados por ‘;’ ou tabulação, por programa social, com

[7] SCE: http://comp-econ.org/.

ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

periodicidade mensal. Cada arquivo possui tamanha variável, tendo em média 1,5Gb de dados, onde sua estrutura interna muda de acordo com o programa social, impactando na leitura e importação dos dados para o ECA.

Todos os dados antes da importação para a plataforma passam por uma normalização, onde são padronizados os formatos dos dados e a sua separação quando necessário, visto que alguns dados são disponibilizados multivalorados.

Na próxima seção será apresentado o trecho de código da plataforma *EconomiC Analyzer*, responsável por realizar a transformação dos dados para o MongoDB.

## *2 ECONOMIC ANALYZER*

A *EconomiC Analyzer* é uma plataforma de coleta, extração e análise de dados econômicos *open source*, que tem por objetivo auxiliar os pesquisadores por meio da centralização de seus estudos sobre programas sociais do Governo Federal em um único local.

Atualmente a plataforma conta com 3 módulos disponíveis, sendo eles: i) *ECA-Importer*; ii) *ECA-Analyzer* e; iii) *ECA-Network*, sua arquitetura pode ser vista na Figura 1.

ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

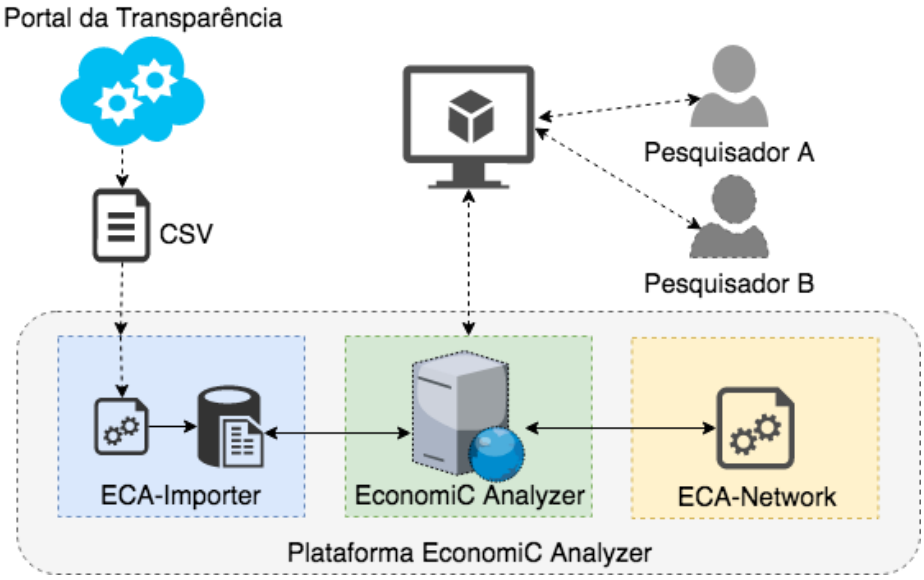


**Figure 1. Arquitetura da plataforma ECA.**

### *2.1 ECA-Importer*

O *ECA-Importer* é responsável por coletar os arquivos CSV, disponibilizados no portal da transparência do Governo Federal, e transmiti-los para o banco de dados. Existem duas versões disponíveis do *ECA-Importer*, uma para banco de dados relacional[8] (MySQL) e a nova para banco de dados NoSQL (MongoDB), disponível em Java[9] e em Python[10] sendo o foco deste trabalho.

No início do projeto o objetivo era desenvolver o importador apenas para banco de dados relacional, utilizando o MySQL[11] como SGBD (Sistema Gerenciador de Banco de Dados) padrão. Contudo, devido ao grande número de arquivos CSV, cada qual com aproximadamente 13 milhões de dados, totalizando de janeiro de 2011 a junho de 2017, mais de 1 bilhão de registros, tornou-se impossível realizar a importação desses dados respeitando as propriedades de um SGBD relacional, em

[8] *ECA-Importer* para MySQL - https://github.com/FIVJ/ECA-Importer
[9] *ECA-Importer* para MongoDB em Java - https://github.com/FIVJ/ECA-Importer-MongoDB
[10] *ECA-Importer* para MongoDB em Python - https://github.com/FIVJ/ECA-Importer-MongoDB-Python
[11] MySQL - https://www.mysql.com/

ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

um banco com as devidas formas normais. Esse problema advém do fato que para importação de cada novo registro deve-se verificar se o mesmo já não existe no banco, evitando a redundância nos dados.

Dessa forma, optou-se pela utilização de um SGBD NoSQL. O termo "NoSQL" significa "*Not only SQL*", onde neste pode-se aplicar o conceito SQL (*Structured Query Language*, ou Linguagem de Consulta Estruturada).

Diferente dos SGBDs relacionais, no NoSQL não há dependência de tabelas e colunas fixas, algo bastante útil nesse projeto, uma vez que os dados disponibilizados pelos programas sociais no portal da transparência são semiestruturados.

Dentre os projetos NoSQL mais notáveis até o momento, podemos citar o projeto de software livre MongoDB[12], que nada mais é que um banco de dados orientado a documentos, o qual armazena dados em coleções de documentos semelhantes a JSON (*JavaScript Object Notation*) compostos por nomes de campos e um tipo específico de valor.

O NoSQL é uma tecnologia que vem ganhando mercado, principalmente quando trata-se de *Big Data*, seja pelo grande volume de dados, falta de estrutura nos dados, descentralidade ou estabilidade dos dados persistidos. Por esses motivos criou-se o *ECA-Importer* para MongoDB.

No *core* do *ECA-Importer* para MongoDB, cada programa social do Governo Federal é uma coleção de dados dentro do banco de dados e todos os dados de cada mês são agrupados, por meio de *mapReduce* são realizados filtros que permitem ponderar os dados dentro de uma mesma coleção e através de agregação dentro do SGBD, é possível vincular os dados que existem entre os programas sociais.

Os detalhes técnicos de implementação do *ECA-Importer* para MongoDB em Java e em Python podem ser visto respectivamente em: i)

[12] MongoDB - https://www.mongodb.com/

ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

https://github.com/FIVJ/ECA-Importer-MongoDB e ii) https://github.com/FIVJ/ECA-Importer-MongoDB-Python.

## CONCLUSÃO

A plataforma *EconomiC Analyzer* começou a ser desenvolvida durante uma iniciação científica, no Instituto Vianna Júnior, e teve como objetivo auxiliar estudantes de pós-graduação em Economia Aplicada, da Universidade Federal de Juiz de Fora, a realizarem seus estudos sobre *social network* no programa Bolsa Família, e como as formações de redes no programa podem influenciar a focalização do mesmo.

O desenvolvimento da plataforma ECA continua ativa e ganhando novas funcionalidade desde 2017. A cada semestre a plataforma agrega novos recursos, novos modelos econômicos e amplia seus horizontes, até então voltados ao campo de estudo econométrico.

Mesmo sendo uma solução relativamente nova, já apresenta grande potencial para a área de economia computacional, visto seus ganhos frente as dificuldades da área de economia em trabalhar com grandes arquivos de dados semiestruturados.

Entre os objetivos futuros, além dos já supracitados, buscamos avançar na análise dos dados disponíveis e na geração de relatórios gráficos e em PDF, extração dos dados em formatos voltados para ferramentas estatísticas e de mineração de dados.

## AGRADECIMENTOS

Os autores agradecem todo apoio e incentivo à pesquisa dados pelo Instituto Vianna Junior e pela Universidade Federal de Juiz de Fora.

ISSN 2176 1035 Jornal ELETRÔNICO V. 10 EDIÇÃO 2 Ago - Dez 2018
Faculdades Integradas Vianna Júnior

## REFERÊNCIAS

BRASIL. Ministério do Desenvolvimento Social e Combate à Fome (MDS). 2016.

Bolsa Família. Disponível em: <http://www.mds.gov.br/bolsafamilia>. Acesso em: 14 de fev. 2018.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, v. 35, n. 2, p. 137-144, 2015.

LOBO, Rodrigo Maciel. O uso de grande volume e variedade de informações-**Big Data**. 2017.

MCAFEE, Andrew et al. Big data: the management revolution. **Harvard business review**, v. 90, n. 10, p. 60-68, 2012.

SIRQUEIRA, Tassio *et al.* Uma Plataforma de Extração e Análise de Dados de Programas Sociais do Governo Brasileiro. **Jornal Eletrônico Faculdades Integradas Vianna Junior**, n. 1, p. 109-128, 2018.

SIRQUEIRA, Tassio; DALPRA, Humberto. NoSQL e a Importância da Engenharia de Software e da Engenharia de Dados para o Big Data. **37º Jornada de Atualização da Informática** (JAI). Congresso da Sociedade Brasileira de Computação (CSBC). Cap. 2. 2018.

SOSA ESCUDERO, Walter. Big data: desafíos para la docencia en Ciencias Económicas. **Econo**, v. 8, 2017.