R-Tableau Connection for Dynamic Big Data Visualization: A Case Study About Brazilian Companies

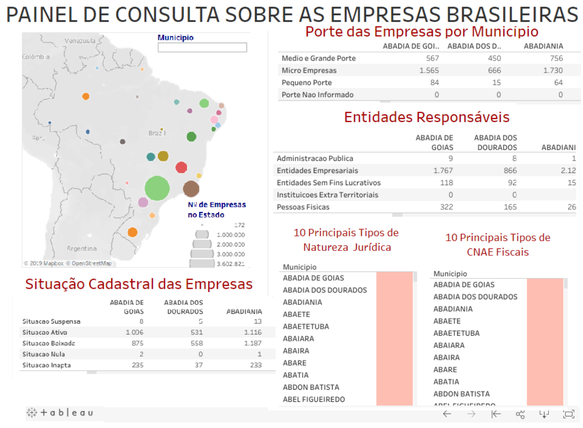
L.L. Boaventura[1] *, A. Ara[1] *;

[1]Universidade Federal da Bahia, BR;

Correspondence to: data.scientist,boaventura@gmail.com

Know that data analytics turns raw data into relevant information. Nowadays, in the age of Big Data, the ability to present this information simply and directly is gaining increasing importance. However, traditional databases simply do not provide the volume and variety of data that comes up every day. That's why companies are turning to technologies such as Hadoop, Spark, and NoSQL databases to meet their rapidly evolving data analysis needs. In this paper, we present one more change in these companies: the R-Tableau connection. Working together, or Tableau has all the exploration, modeling, and data processing capabilities, do R, as well as all the simplicity, lightness, and iteration Tableau has.

To illustrate this connection, in addition to the theoretical context, we use mass data analysis and processing techniques to extract information with respect to the available CNPJ database, as determined by the 1988 Federal Constitution c / ca Access to Information Act. c / co Federal Decree No. 8.777 / 2016, on the website of the IRS. This data represents information about all companies registered in Brazil, and is therefore one of the most relevant sources of public information in the country. However, even though the CNPJ database is available for download, the IRS publishes a *fixed width - fixed width file* format, or makes it impossible to directly use software such as R and Python for parsing data. Dice. Added to that, the unzipped files are over 85GB, or further limiting broad access to the general public. Thus, in this work, uses the Sotware R Core Team 2019, through the library *qsacnpj*, developed by the Social Observatory of Brazil, located in Santo Antonio de Jesus, Bahia State, to change the CNPJ data *fixed width - fixed width file* format in csv. So that after that, through the *sparklyr* and *ggplot2* libraries, Sotware R also analyzes the data and extracts useful information for society, a presentation purpose, with R and *Tableau*, through a visual and dynamic dashboard, making CNPJ data much more interpretable.



R-Tableau Connection

Tableau is an American company, a leader in the development of data visualization software. Tableau Software is a business intelligence program that enables the creation of Interactive Dashboards, enabling the user to transform data into information that helps make more informed decisions. All in all, Tableau is a complete and easy-to-use analytics platform as it provides complex drag-and-drop graphing.

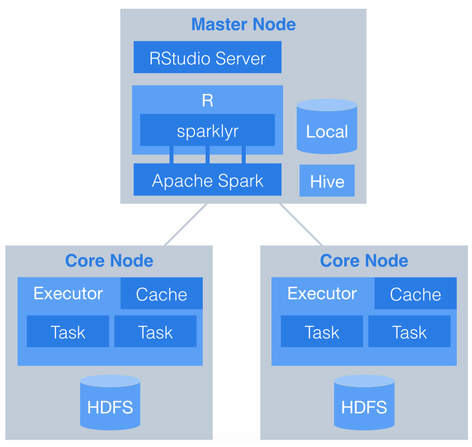
As mentioned earlier, R is a respected open source environment for statistical computing and graphics output, in which there are thousands of users who constantly add new coding content and practices. Thus, starting with Tableau version 8.1+, these users can integrate R encodings into Tableau's calculated fields. This integration of Tableau with R enables the user to send Tableau information to the R engine, perform data calculations, and send the information back to Tableau.

For the connection to be effective, users must have access to an R server to access R functions in Tableau Desktop or Tableau Server. Tableau needs to communicate with R revolution using an R package called ' textit Rserve'. First, you must reread your installation on R to start Rserve on Tableau. Rserve is located in Tabeleau's 'Help' menu, where the user must choose the 'Manage R Connection' option. Then just hit "ok" and the templates as well as any R packages or functions that are installed on the server can now be used in Tableau.

Big Data in R by Sparklyr

Before talking about sparklyr it is important to understand what Apache spark is. Apache Spark is a unified analysis engine for large-scale data processing. That is, Apache Spark is a big data tool that aims to process large datasets in a parallel and distributed manner. So textit sparklyr is a package that provides an interface between R and Apache Spark. It arose from the need for R users to have a native dplyr interface for Spark.

Thus, textit sparklyr lets you parse data in Spark while you use familiar tools in R. Because it supports a full dplyr backend, a popular tool for working with in-memory and in-memory data frame objects insufficient, sparklyr enables the user to use dplyr to convert R code to Spark SQL. In addition, textit sparklyr supports MLlib, making it possible to run classifiers, regressions, clustering, decision trees, and many other machine learning algorithms on your distributed data in Spark. In this work, specifically, we use textit sparklyr to filter and aggregate CNPJ datasets into Spark, then bring them to R for analysis and visualization.



Case Study: CNPJ Dataset

CNPJ's public data refer to the registration information of all companies registered in Brazil. This data is collected on a quarterly basis and is available for download as a fixed width file from the IRS website. The CNPJ database is divided into the site into 20 parts, denoted by: Open Data CNPJ n, where n = 1, 2, ..., 20. For the analysis of CNPJ data, some attachments containing technical terms and codings used to database construction, have their required prior reading. These are: *layout* and Cadastral Situation Reason Table.

The attachment *layout* describes the codes used for each of the attributes of the bank variables, divided into three major groups: HEADER, which contains the information about how the record was generated; MAIN LAYOUT, which contains the registered registration items of each company; and LAYOUT SOCIOS, which comprises information about the legal guardians of each company.

In this paper, we use the information contained in the MAIN LAYOUT only. Arranged in the $cnpj_data_cadastrais_pj.csv$ file resulting from the application of the *qsacnpj* package described in Section 2.1.1 of this document contains 40 variables about information specific to each registered company in Brazil.

CNPJ Card Example: In Brazil, the National Register of Legal Entities (acronym: CNPJ) is a unique number that identifies a legal entity and other types of legal arrangement without legal personality (such as condominiums, public agencies, funds) with the Brazilian Internal Revenue Service (Ministry of Finance). from the farm). The CNPJ Card comprises all registration information of the entities of interest of the tax administrations of the Union, the States, the Federal District and the Municipalities and is required for processing (Art. 15 of Law 11,419 / 2006).

Web Scraping and Transformation of CNPJ Dataset

The Library *qsacnpj* was developed and executed by the Social Observatory of Brazil, located in the municipality of Santo Antonio de Jesus, in the state of Bahia. qsacnpj was developed to be an ancillary tool for the Public Account Transparency project.

The function of *qsacnpj* is to make the CNPJ database more accessible to society by placing it in a readable standard for most machines. This is necessary because, as mentioned earlier in this paper, the IRS makes the CNPJ bank available in a fixed text format, which requires a high degree of data processing expertise, so that the bank is then analyzed.

Thus, George Santiago, employee of the Social Observatory of Brazil, and creator and developer of *qsacnpj*, suggests that the following steps be taken to apply the package:

1. Download the (.zip) files from the CNPJ database from the IRS website.
2. Save the '.zip file to the directory that will be used for data processing.
3. Unzip the files in the directory on a machine that has sufficient space. Given that the unzipped files are over 85Gb.
4. Add the extension '.txt' to the end of the file name.
5. Create a specific folder to store only the '.txt' files.
6. Install and run the package *qsacnpj*.

Within the *qsacnpj* library, there is the $generate_bd_cnpj()$ function. In it, it is possible to transform the data to both SQLite and CSV formats. For the CSV format used in this work, simply define within the $generate_bd_cnpj()$ function, the name of the saved file ($path_files_txt$), the number of lines handled at each iteration ($n\ ines$) is the format to store the transformed file ($store$), in this case ($store = "csv"$).

This done, at the end of the processing, the user will see, in the chosen directory, the CSV files: ($cnpj_data_contact_pj.csv$, $cnpj_data_pj.csv$ and $cnpj_data_cnae_secondary_pj.csv$).

Dashboard Build Process

Of the 40 variables in dataset, only the following were considered: "data_situacao_cadastral", "bairro", "municipio", "motivo_situacao_cadastral", "codigo_natureza_juridica", "uf", "cnae_fiscal", "porte_empresa", "data_inicio_atividade", "capital_social_empresa" e "situacao_cadastral". As a result, in a database with 11 variables, presented through 40 million 613 thousand 847 companies.

CNPJ Dataset Dashboard Build Cycle: R Studio) Download R Studio and a package *qsacnpj*. **qsacnpj)** Use this library for webscraping CNPJ data. **sparklyr)** Use this library to work on R with big data. **Rserve)** Use it to connect R to Tableau. **Tableau)**Build the graphs and filters to visualize the treatment done.

The database was divided into six query factors: Geolocation, Size, Responsible Entities, Legal Nature, CNAE Fiscal and Cadastral Status. The objective then becomes to extract as much information as possible from the companies in each municipality. For this, each of the six factors would be presented separately within the same panel.

The Geolocation factor, presented by means of a map, would be responsible for providing the user with a visual comparative analysis of both municipalities and states. The factor Porte, makes available to the user, to analyze the amount of companies in each of the municipalities, separated by size. Similarly, the Registration Status and Responsible Entity factors allow the user rapid comparisons and conclusions, since there is no data loss for these two factors. Different situation of factors Legal Nature and Fiscal CNAE. Due to the great diversity available for both types of CNAE and types of Legal Nature (both with more than 50 possibilities), we seek to present only the 10 most frequent observed in each of them.

Consult the Dashboard

To consult the Dashboard resulting from this work just access the QRCODE available in the upper right corner.

References

[1] Chaomei Chen. Top 10 unsolved information visualization problems. *IEEE computer graphics and applications*, 25(4):12–16, 2005.

[2] Andrea De Mauro, Marco Greco, and Michele Grimaldi. What is big data? a consensual definition and a review of key research topics. In *AIP conference proceedings*, volume 1644, pages 97–104. AIP, 2015.

[3] Kawa Nazemi. Semantics visualization. In *Adaptive Semantics Visualization*, pages 69–121. Springer, 2016.