DATA PIPELINE: IMPROVING MANAGEMENT OF FINANCIAL CONTRIBUTIONS TO THE FIGHT AGAINST POVERTY IN COSTA RICA

Topics: Social Sciences and Databases/Data Management

Author: Roberto Delgado Castro
Affiliation: Dirección General de Desarrollo Social y Asignaciones Familiares (DESAF), Ministry of Labor and Social Security (MTSS), Government of the Republic of Costa Rica.

ABSTRACT

FODESAF, administered by DESAF (part of the Ministry of Labor and Social Security), is Costa Rica´s and Latin America´s largest public-social investment fund. It transfers around US$1.000 million per year (2% of local GDP) to a wide variety of social programs natiowide.

Local employers (patrons) and Ministry of Treasury (Government) provides its economic resources due to monthly financial contributions.

Among with monthly financial transfers, a large database with specific information of contributors is attached. Since 1978, FODESAF´s establishment year, DESAF have not had the opportunity to classify and analyze such crucial data.

A data science project was developed in SQL© and RStudio© to implement a *Data Pipeline*, in which all data was loaded to break down its key elements, in order to improve DESAF authority´s decision-making capabilites.

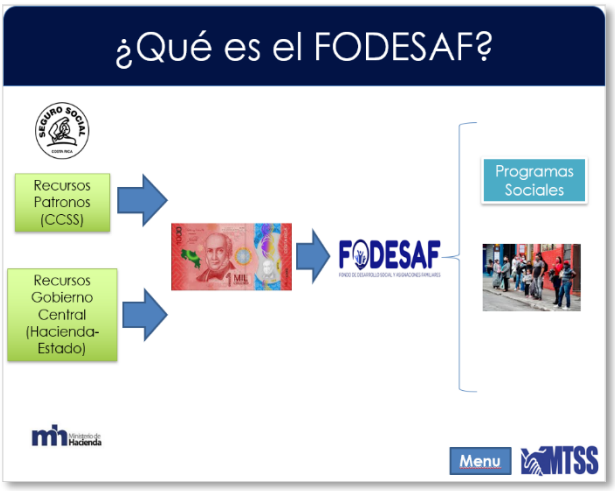
As inputs, annual databases within 2003 and 2019 (17 years) were loaded into the Pipeline in a separate way (250.000 registers per year, 30.8 million in total). The results were RMarkdown© automatic reports with brand-new dataframes and visualizations for each year, that helped authorities visualize and analyze elements that had not been seen in 43 years of DESAF´s history.

After its implementation, local government has now a unique-recurrent data science tool to improve management of financial contributions to the fight against poverty. As key learnings, data-taming skills were strenghten, project-questions were defined as project´s coding-structure strategy, brand-new *Contributors Mass Report* was developed, and this project has been used as an economic-recovery follow-up-instrument in post-pandemic era.

Keywords: Data Pipepline, data-taming, visualizations, automatic-reports, poverty.

What is FODESAF?

FODESAF (Fondo de Desarrollo Social y Asignaciones Familiares), administered by DESAF (Dirección General de Desarrollo Social y Asignaciones Familiares) is the main social-public investment fund in the fight against poverty in Costa Rica. It's a Latin America´s unique, supportive and law-fixed-guaranteed financial fund that transfers around US$ 1.000 million per year (2% of local GDP) to 20 institutions that executes, in turn, 25 social programs nationwide.

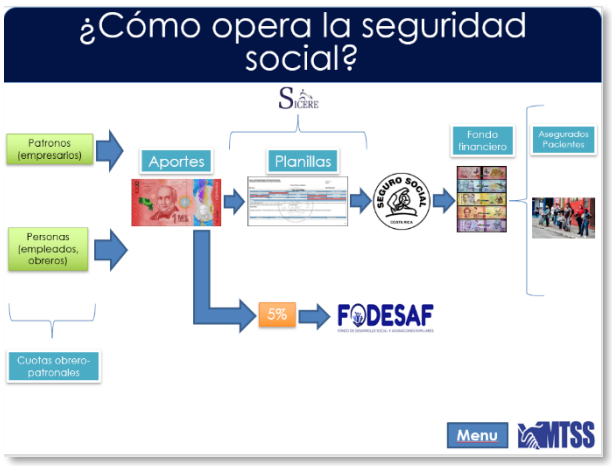


Its financial resources come from law-established contributions of patrons, which are employers and companies (60%) and from National Budget of Treasury Ministry (40%). Contributions from patrons are collected, admninistered and transferred to FODESAF (in a monthly basis) by Caja Costarricense de Seguro Social (CCSS), the guiding institution of the social security sector in the country.

Since then, FODESAF is the most important public-policy intrument in the fight against poverty in Costa Rica.

How social security operates in Costa Rica?

All local employees and their employers (patrons) contributes with economic resources, in monthly basis, to constitute a large financial fund for social and medical assitance. Employees contributes with a percentage of their income (salary), and employers contributes with a percentage of their revenue. Such contributions are called "Cuotas Obrero-Patronales".



Such contributions from both participants are registered in payroll forms and sent to Caja Costarricense de Seguro Social (CCSS) the leading social-security institution which main task consists in the management of the mentioned social-assistance large fund. Final benefitiaries of such fund are all contributing-citizenship any time they need medical assistance.

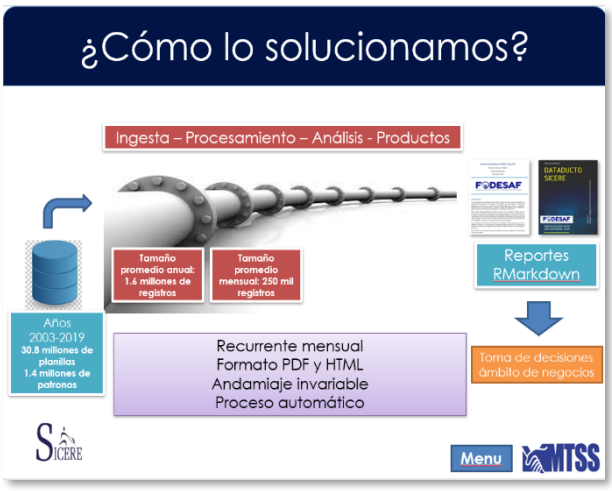
All payroll forms logistics are executed by a CCSS´s agency called SICERE (Sistema Centralizado de Recaudación). In fact, such agency is the one in charge of sending to DESAF the large database among with the monthly financial transfers from contributors.

Problem confronted

The typical-historic logistic-operative path that DESAF has been following since 1978 with contributions from employees and employers (yearly basis), consists only in receiving the economic resources and, consequently, include such amounts in budgets in order to define specific expenses plans. Among with such transfers, DESAF also has received large databases from SICERE with specific information of the correspondent contributors.

The problem resides in the fact that DESAF have not had the opportunity to classify and analyze such data, in order to recognize specific details and crucial information of contributors, such as the quantity, type and up-to-date conditions of payrolls, economic activity, quantity and geographic locations of employers and employees and other key specific classifications. Due to that situation, crucial data had not been analyzed properly, and highly-relevant insights have not been shown in order to improve public-policy-quality of DESAF´s and government´s authorities in the fight against poverty in Costa Rica.

Problem´s solution

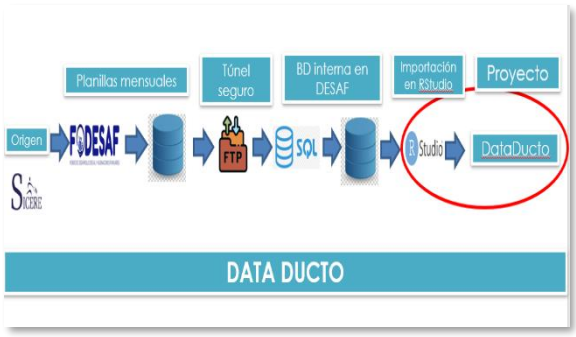


To solve the mentioned problem, DESAF has designed a *Data Pipeline* in SQL© and RStudio©, in which all data from SICERE was loaded to break down its details and key elements, in order to improve DESAF authority´s decision-making capabilites.

As inputs, annual databases within 2003 and 2019 (17 years) were loaded into the Pipeline in a separate way (17 databases, 250.000 registers per year, 30.8 million in total). The results were RMarkdown© automatic reports with brand-new dataframes and visualizations for each

year, that helped authorities visualize and analyze elements that had not been seen in 43 years of DESAF´s history.

The process begins with the data extraction from a secured FTP (File Transfer Protocol), in which SICERE puts the databases. Due to SQL codes, the databases are extracted from FTP and storaged into DESAF´s local secured-virtualized servers. Later on, such databases are imported into RStudio to load them into the Pipeline (Data Science project).



Key learnings

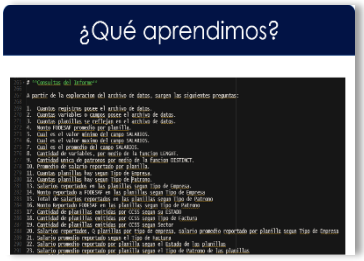
The design of the mentioned Pipeline generated key learnings, as follows:

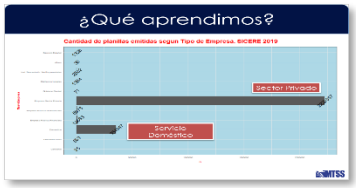
1. *Data-taming* of variables from the original database received from SICERE. We performed functions from Tidyr© and Dplyr© packages to order information from raw databases. Also, we used Lubridate© package functions to adjut date´s formats.

2. Prior to execute the coding of the Pipeline, we defined *project-questions*, which were the specific information that we needed to extract from the databases. Such questions were crucial in order to set up the specific path to follow with the coding issues. This task saved us a lot of time with coding and helped us not to get lost into the large amount of variables of the original database received.

3. *Brand-new visualizations* using ggplot© package generated high-quality RMarkdown reports.

4. *Automatic reports* (results) were developed using RMarkdown© package. Such reports were displayed in HTML and PDF formats, either for sending them by email to DESAF´s authorities, or publishing them in FODESAF´s website.





5. A brand-new *Contributors Mass Report* was developed in order to identify, analyze and evaluate the evolution of FODESAF´s contributors in time.

6. The analysis of the evolution of financial transfers of contributions from employers and employees, has been utilized by DESAF´s authorities as an instrument to evaluate and determine *domestic economic recovery* during the post-pandemic era.

7. In general terms, local government has now a unique-recurrent data science tool to improve management of financial contributions to the fight against poverty. Additionally, we have proven that a *data-science project* can be implemented succesfully in Costa Rica´s public sector using RStudio©.

Special indication

I presented this data science project during *ConectaR Conference 2021*, in the Civic Science topic. January 28th and 29th, February 4th and 5th, 2021. San José, Costa Rica. This research has not been published anywhere.