Enhancing Data Quality: A DataOps approach with R and GitLab

Alexandre Cunha, Bruno Lima, João Poças

Statistics Portugal

Abstract

DataOps has emerged as a fundamental practice, essential for optimizing management and delivery of data-products. Moreover, this framework is designed to enhance collaboration between those involved in the data lifecycle. By the same token, using R for data analysis facilitates the implementation of streamlined and reproducible data workflows. Its integration with GitLab, a version control and collaboration platform, empowers teams to efficiently manage and version data projects, fostering reproducibility and ensuring that analyses can be recreated and validated by other team members.

The convergence of *DataOps*, R, GitLab, reproducibility, and data pipelines represents a powerful paradigm shift in data science and in the statistical processes. Here, we show some insights into how these elements can be effectively combined to foster collaboration, increase the efficiency of data workflows, and ensure the reproducibility of results in the dynamic and fast-paced world of data-driven decision-making.

Keywords: DataOps, Data-driven, R, Collaboration, Process management

1. Introduction

The implementation of a *DataOps* methodology centred on the R programming environment enables the development of data pipelines guaranteed by high quality, automated and reproducible processes. Bearing in mind the principles of the *DataOps* Manifesto (The DataOps Manifesto,2023), the aim is to design processes that meet these principles and facilitate the delivery of data products. The *DataOps* approach presents a more efficient way of managing data in an organisation. Using a set of practices, methods, and technologies, it is possible to achieve gains in speed and collaboration, while promoting a culture of continuous improvement in data processing.

The R package {targets} (Landau, 2021) is a key tool for the definition of data pipelines that help to streamline the movement of data from source to destination. The integration of data pipelines with R and GitLab facilitates the automation of data extraction, transformation, and loading (ETL) processes, reducing manual effort and enhancing data quality.

Using GitLab facilities like “Issues” and “Issue Boards” in a *DataOps* context can significantly boost project management, collaboration, and tracking within data science workflows. GitLab Issues provide a structured way to define, assign, and track tasks or objectives within data projects, where each issue can represent a specific piece of work, such as data cleaning,

model development, or report generation. Meanwhile Issue Boards provide an at-a-glance view of project progress. Likewise, team members, stakeholders, and managers can quickly see which tasks are pending, in progress, or completed, promoting transparency.
In Statistics Portugal, every month, data is collected on each trade exchange through business surveys and administrative data sources. These data reflect imports and exports of goods and services and are used to characterise Portugal's trade flows with other countries. Either from manual recording or from administrative sources, data may contain errors and/or anomalies that need to be identified and corrected. Statistics Portugal, as the national authority for collecting, processing and disseminating statistical data, is responsible for complying with the regulations and requirements imposed by the EU in order to provide data on trade involving Portugal. The quality of these data is monitored by the International Trade Data Collection Unit (ITDCU) as soon as they are received by Statistics Portugal.

1.1 Aim

Describe the process of developing and implementing an application to identify anomalies in data on extra-EU transactions (imports and exports) analyzed by the ITDCU, based on a DataOps methodology.

2. Methods

In order to efficiently organise the tasks to be carried out in collaborative work, GitLab issues were used to record these tasks. The use of an Issue Board to organise the issues created provided a visualisation of the workflow where the status of all tasks could be seen. With the issue board it was possible to easily follow the progress of this project and track it to ensure that nothing was forgotten.

Initially, two main groups of tasks were defined: a) the preparation of tables and views in Oracle-based RDBMS with the original data to be processed; b) the construction of an R package that would compile the functions to be applied to the data to obtain an anomaly score for the data received.

The R package {mdaitr} (Lima & Cunha, 2024), created explicitly for this project, has been developed using functional programming logic, which allows not only the creation of documentation, and testing of each function but also their application in an interconnected workflow, with more concise and readable code. The flexibility of the R language allows functional code to be parallelized and large problems to be solved with few changes, reducing the risk of error.

A data processing pipeline has been defined using the {targets} package (Landau, 2021). This package allows reproducible workflows to be maintained in pipelines whose objects are only executed when necessary, ensuring more efficient data processing. It checks for dependencies in the workflow, only running the outdated parts, with a clear benefit in terms of workflow execution time.

The defined pipeline translates into an efficient, reproducible ETL process for the data to be processed, based on functional programming. The data is originally loaded from views created in Oracle, analysed, and transformed using the {mdaitr} package and finally the processed data can be read by the {shiny} application (Chang et al., 2023) to which end users have access.

The history of changes to this project's source code was maintained using Git, a free and open-source version control system. The use of GitLab enabled the creation of work repositories in the cloud, mirroring the local repositories and allowing all work to be organised and shared. GitLab as a source code hosting platform allows different members of the working team to contribute to the project and also to validate and review the code before it is merged with the main code.

3. Results

The {mdaitr} (Micro Data Analysis on International TRade) package (Lima & Cunha, 2024) compiles the functions needed to process the data to apply the so-called 'scores model', which validates the values of the variables being analysed. This scoring model was adapted from an implementation of the Swedish Foreign Trade Statistics (SFTS) (Norberg & Jader, 2005). For each answer, a score is obtained which translates the possibility of it being an anomaly, considering the suspicion calculation and the impact of this value on the total per product. Briefly, the model in question uses a *score()* function calculated as a weight of geometric averages of suspicion error and potential impact on the information to be disclosed.

To analyse the data and identify suspicions, an automated process is defined in a pipeline that establishes the data flow, using the {targets} package (Landau, 2021). This ETL process is carried out periodically according to a schedule predefined by the ITDCU. Thus, the workflow can be categorised into data extraction, data preparation, computations (including the scores), and data output.

Using the {targets} package (Landau, 2021) makes it possible to maintain a reproducible workflow and avoid repeating tasks. As {targets} recognises tasks that are already up to date, processing time is optimised, and tasks can also be executed more efficiently since this

package supports implicit parallelisation. The files are treated as R objects, which makes the data easier to manipulate and manage, and provides tangible evidence that the results match the underlying code and data.

An application developed in Shiny (Chang *et al*. 2023) provides the results of data processing in the workflow. In addition to an overview of the analysed observations, it is also possible to have access to the history by product code for the value and quantity variables.

4. Conclusions

A *DataOps* approach makes it possible to define more efficient workflows for analysing data, from its origin to the final product. Based on agile methodology, the aim is to use the best practices, processes, and technologies available for data management.

The definition of automated statistical and analytical processes (Reproducible Analytical Pipelines (RAPs)) is based on best practices in software engineering, with the aim of ensuring data pipelines that are reproducible, auditable, efficient and of high quality (GAP, 2024).

Anything that can be automated should be automated, so the choice of functional programming will reduce the repetition of code and thus minimise the potential for error. Traceability and reliability are essential pillars for the success of *DataOps* projects (Rodrigues, 2023), ensuring that data is handled efficiently, transparently, and reliably. Even though there are no specific tools for building a RAP, using the R programming environment gives us the ability and flexibility to define a process from the source of the data to the final product. However, automating statistical production with RAP implementation can be challenging, especially when it requires changing practices that are already ingrained in the culture of institutions (Gregory & Upson, 2019).

In conclusion, the convergence of *DataOps,* R, GitLab, reproducibility and data pipelines represent a powerful paradigm shift in data science and the production of official statistics. Here, we present some concrete examples on how these elements can be effectively combined to promote collaboration, increase the efficiency of data workflows, and guarantee the reproducibility of results in the dynamic and fast-paced world of data-driven decision-making.

References

Government Analysis Function (GAF). 2024. Reproducible Analytical Pipelines (RAP). https://analysisfunction.civilservice.gov.uk/support/reproducible-analytical-pipelines/.

Gregory, M & Upson, M. (2019). Reproducible Analytical Pipelines. UKgov. https://ukgovdatascience.github.io/rap_companion/.

Landau, William Michael. (2021). The Targets r Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing. Journal of Open Source Software 6 (57): 2959. https://doi.org/10.21105/joss.02959.

Lima, B., & Cunha, A. (2024). Micro Data Analysis on International TRade. In-house repository.

Norberg, A & Jader, A. (2005). A Selective Editing Method Considering Both Suspicion and Potencial Impact, Developed and Applied to the Swedish Foreign Trade Statistics. UNECE. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2005/05/sde/wp.12.e.pdf

Rodrigues, Bruno. 2023. Building Reproducible Analytical Pipelines with R. amazon. https://raps-with-r.dev/.

The DataOps Manifesto. (2023). https://dataopsmanifesto.org/en/.

Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. (2023). Shiny: Web Application Framework for r. https://CRAN.R-project.org/package=shiny.