
INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly
financed by the European Union
CASINO ESTORIL
HOTEL PALÁCIO
EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Enhancing Data Quality:

A DataOps approach with R and GitLab

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Summary

1. Motivation
2. Data Pipelines in International Trade Data Collection
3. GitLab as a DataOps board
4. Improving data quality with R packages
5. Conclusions

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Motivation

International Trade Data Collection Unit (ITDCU)

- INTRA: imports/exports with EU members
- EXTRA: imports/exports with non-EU members

Daily Routine

- Data is updated twice a day
- May contain records from up to a year ago

Score Model

- Implementation of the Swedish Foreign Trade Statistics (SFTS) *(Norberg & Jader, 2005)*
- Considers not only the suspicious error but also the potencial impact of the record

Microdata Analysis

- ITDCU perform daily analysis
- Suspicious get ranked and distributed to colleagues

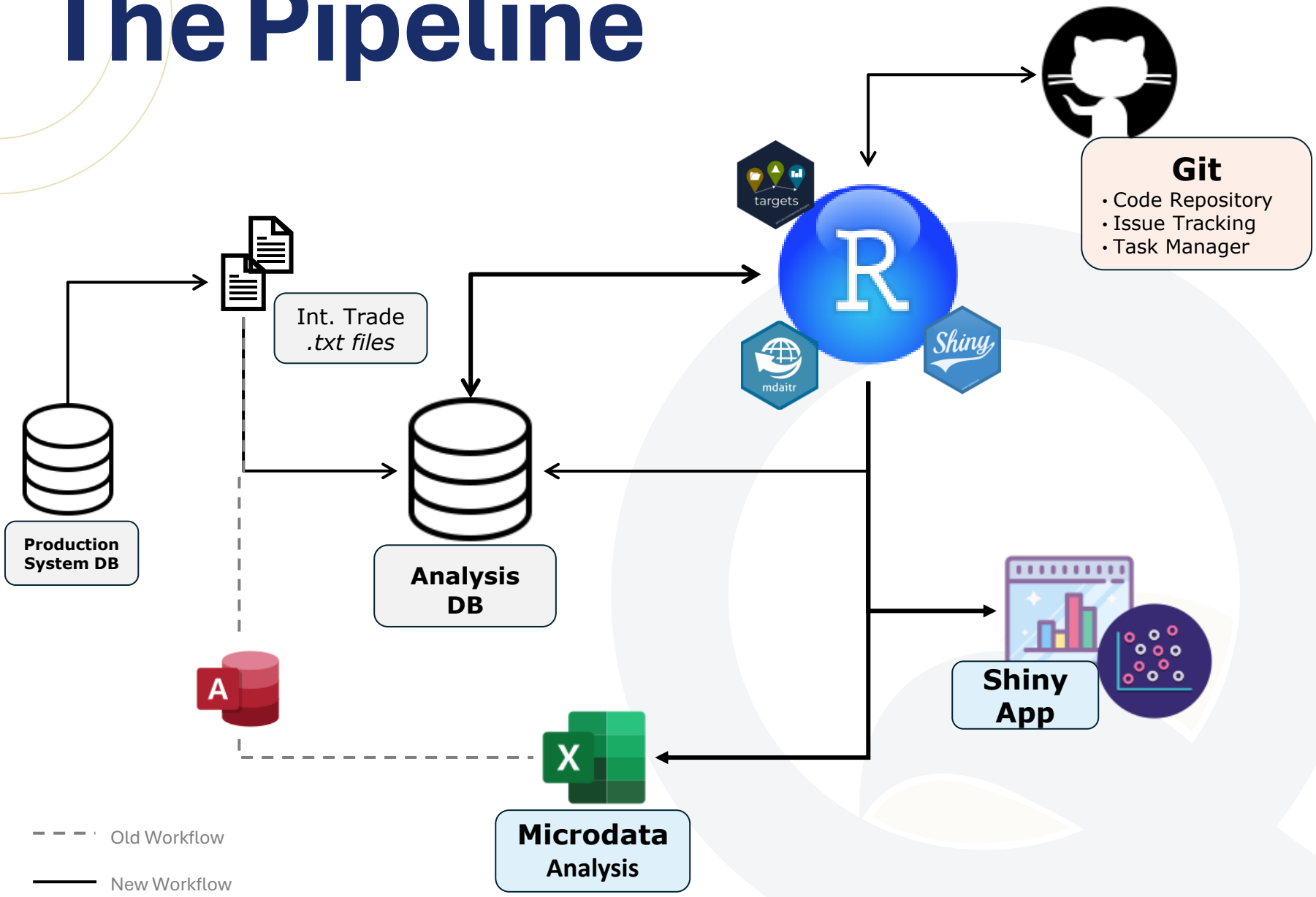
INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

The Pipeline



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

DataOps

CI/CD approach

- High quality, automated, reproducible processes
- Culture of continuous improvement in data processing
- Speed and collaboration

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

GitLab

DataOps Board

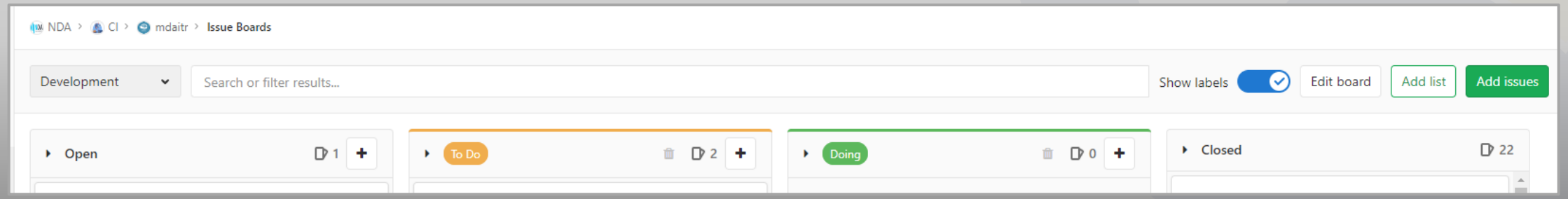
- Define, assign and track tasks or objectives
- At-a-glance view of project progress
- Issue Tracking and notifications
- Version Control

The conference is partly financed by the European Union

GitLab

DataOps Board

- Define, assign, and track tasks or objectives
- At-a-glance view of project progress
- Issue Tracking and notifications
- Version Control



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

R

ETL Workflow

Load data from RDBMS databases

Calculate the *score value*

Export to .xlsx

Update Shiny App

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

R

ETL Workflow

Load data from RDBMS databases

Calculate the *score value*

- Calculate quantiles for each Combined Nomenclature
- Calculate Suspicious Error
- Calculate Potential Impact
- Calculate *score()*

Export to .xlsx

Update Shiny App

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

R

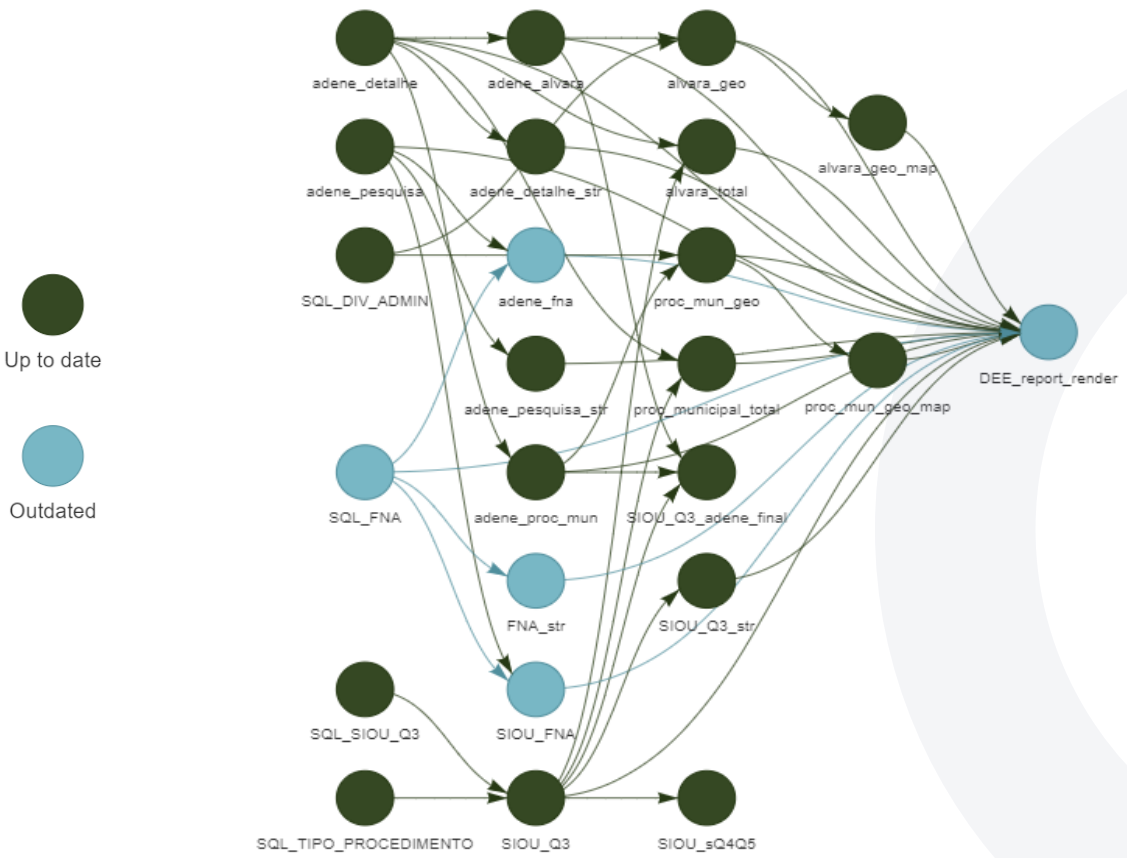
{mdaitR}

- Adapted from an implementation of the Swedish Foreign Trade Statistics (SFTS) *(Norberg & Jader, 2005)*
- Developed under the Functional Programming paradigm
- Creates tested and documented functions
- Can be applied to other data domains

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

R

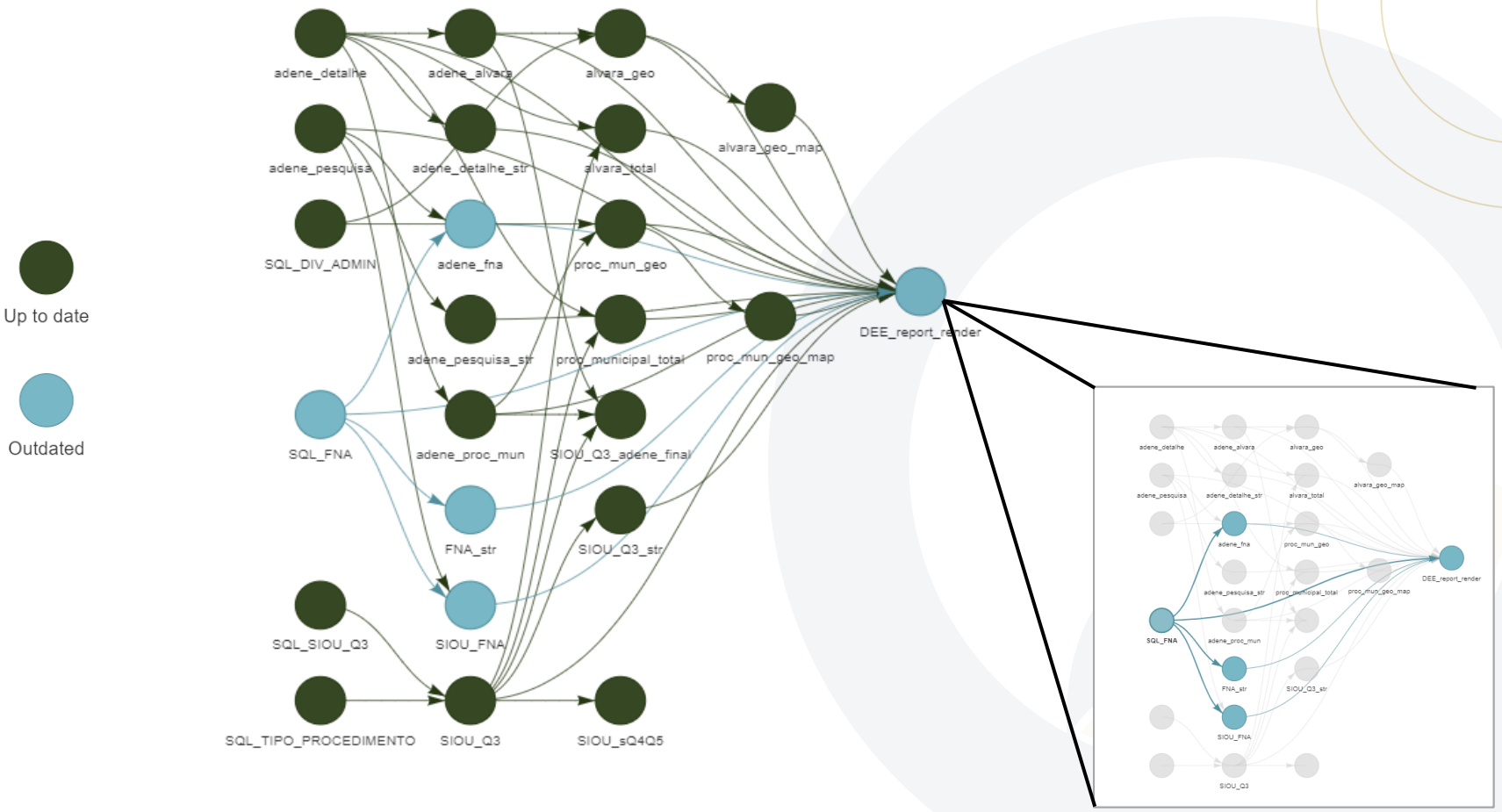
{targets}



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

R

{targets}



EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Conclusions

- RAPs are based on best practices with the aim of ensuring data pipelines that are reproducible, auditable, efficient and of high quality
- Anything that can be automated should be automated!

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Thank You!

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union