User empowerment with new technologies

DataOps journey on SPACE platform at ECB
What is DataOps?

DataOps focuses on automation, flexible data science platform and promotes communication between, and integration of, formerly siloed data, teams, and systems.
System Architecture and Services (SPACE)

Business-specific code
Common Functions Library

Data Access Layer (DAL) {REST:API}

Data stages
Collect -> Rec
Prod -> Compile
Pre-Diss -> Diss

Metadata
# Data production (value pipeline)

## Camunda
- Automation, orchestration and monitoring
- Is used for executing the standard SDMX data production (validations, aggregations, reporting, dissemination etc.)
- Accessed either directly or via Web Portal
- Users have read-only access to Camunda and DAL system logs

## JupyterHub
- Used for direct access to data / ad-hoc data analysis
- Users develop and test in JH. Deploy mature code to Camunda
- All production steps can also execute via JH
- Execution log visible in real time

## Rshiny via Posit Connect
- Used to publish dashboards
User self-service (innovation pipeline)

**Deployment**
- Centrally maintained common components
- Users maintain the business-specific code (Python and R)
- Users can also modify (and later deploy) Camunda workflow definitions
- Changes are pushed / merged to GitLab
- Deployment is via GitLab click to any environment
- Users can configure process metadata and SDMX metadata (constraints)
- Challenge: ensure appropriate access rights

**Monitoring**
- Workflow monitoring via Camunda or directly in JH
- Access to Camunda execution logs and to system logs

**Versioning**
- Data
- Metadata
- Business process configurations
Automated GitLab pipelines

Innovation Pipeline

• GitLab is repository for all CFL, business code and configurations
• Users maintain their own pipelines in GitLab
• Pipelines are used for functional and performance regression testing
• Triggered manually or on schedule (to test before new code deployments for example)
• Choose which environment to run in
• Multiple Python notebooks can be created and executed sequentially
• Ideally notebooks also generate the test data needed for the test
• Pipeline execution log viewable
Data integration eases innovation

### Data sources

- SPACE HBase: R/W
- ECB corporate datastore (Impala / Hive): R/W
- FAME legacy system: R/W
- ECB Statistical Data Warehouse (SDW): R

### Data Integration

- Exposed to users in a uniform way via Python / Pandas dataframes (in memory)
- At low level data is mapped in a unified data model in JSON
- Accessible via REST API for integration with third-party tools
Conclusions

• Implementation of DataOps on SPACE is an ongoing, evolving effort
• It empowers users with better quality control over data and code
• While users maintain their codebase themselves, they still require support with guidance and best practices for implementation
• Transitioning to the new Python / R based system, also means the userbase of economist and statisticians is developing a new role as data scientists
• To steer the development of the common part of the library, a user community is needed
• From IT perspective: requires some changes but tools are shared with DevOps