ATINER CONFERENCE PRESENTATION SERIES No: COM2022-0248

ATINER's Conference Paper Proceedings Series
COM2022-0248
Athens, 9 September 2022

Creating Modern Data Lake Automated Workloads for Big Environmental Projects

Mihail Mateev

Athens Institute for Education and Research
9 Chalkokondili Street, 10677 Athens, Greece

ATINER's Conference Paper Proceedings Series
COM2022-0248
Athens, 9 September 2022
ISSN: 2529-167X

Mihail Mateev, Chief Assistant Professor, University of Architecture, Civil Engineering and Geodesy, Bulgaria

Creating Modern Data Lake Automated Workloads for Big Environmental Projects

ABSTRACT

Data Lakes provide a modern approach to persist data with heterogenous structure for different types of analysis. It offers centralized repository that allows to store all structured and unstructured data at any scale. Big environmental projects nowadays include data from different sources, that need to be approved, managed, processed and later shared for specific analysis. Data Lake automation flows offer one modern approach to manage data in similar solutions. This technology can be used from any kind of big organization: government, non-profit or Commercial. This paper demonstrates a methodology how to build modern repository and implement automation flows to approve, process, analyze and share data for big environmental projects. In the research are included real life cases, demonstrated with prototypes using Azure Data Lake and automated flows with Microsoft Power Platform and Azure Data Factory. Automation flows include cloud workloads as well as Robotic process automation (RPA) flows that enables engineers and non-coders alike to automate processes and tasks across desktop and web applications.

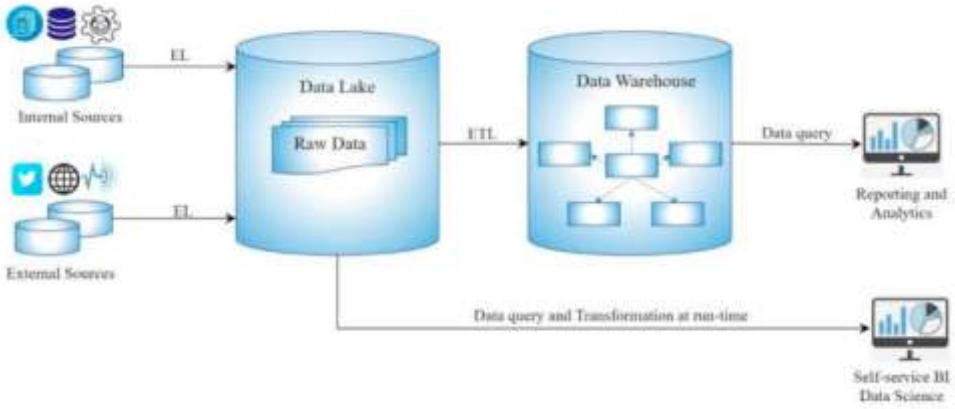
Keywords: automation, automated flows, big data, machine learning, data lake, data management, environmental engineering, Microsoft Azure, Microsoft Power Platform, robotic process automation

Acknowledgments: Special thanks to my family for the support and Microsoft for the Azure resources, which made possible this research.

Introduction

There are three purposes of data lakes: as staging areas or sources for data warehouses, as a platform for experimentation for data scientists or analysts, and as direct source for self-service BI, as illustrated in Figure 1 [9].

Figure 1. *Different Purposes of Data Lakes*



Source: Llave, M. (2018) "Data lakes in business intelligence: reporting from the trenches", Procedia Computer Science [9].

Automation is a critical part to have effectiveness using data lake solutions. This paper has been developed to describe at a general level the data lake and workload automation approach for environmental data used in data management for the implementation of the Flood Directive in Romania, realized as a big data project.

The project “Flood Hazard and Flood Risk Maps and Flood Risk Management Plans for Romania” has a significant degree of complexity due to the diversity in typology of activities, starting from survey works, continuing with hydrological and hydraulic modelling in a variety of hypothesis and situations and finishing with the study of socio-economic impact of the measures applied for flood risk mitigation. Other contributing factors are the existence of multiple stakeholders involved, the national nature of the project, and the volume and variety of information involved. Therefore, it is essential to have a reliable data management system. A clear and coherent information structure will contribute to the proper development of the project in two main directions:

- Easy access to information
- Clarity regarding the storage and naming of data for all distribution files created during the current FD implementation cycle.

Overview of the Data Organization and Data Structure Development

There are two important concepts that were considered in development of data structure for FD Implementation:

- grouping and organizing the information into a coherent structure.
- naming folders and files.

Project has 2 cycles, described in the methodology part –

- Cycle 1, where part of data is already collected into a specific structure
- Cycle 2, including data models for the hazard risk and organized in the concept collected data to be used from different consumers.

To the data structure for 2nd Cycle of FD Implementation was for defining the data structure for Cycle 2 Implementation of FD, an analysis of the data generated in Cycle 1 was accomplished. In this regard, WB requested and collected all data associated with the projects from Cycle 1 at national and at RBA level, including data for Danube River. It is important to mention that each RBA plus Danube UoM, managed their implementation individually, therefore there are 12 different manners of implementing Cycle 1.

After an analysis of the data received in association with Cycle 1, two main processes were made:

- Separating the data which was generated within the project's scope (deliverables) and the data which was used for developing the technical products in Cycle 1 (which played the role of input in the project).
- Reorganizing the deliverables from Cycle 1 using a unique and organized structure at macro level.

Big Data Projects Specifics

Big data projects now have the following specifics:

- Contain unstructured and structured data
- Data is partitioned
- Have schema-on-read semantics
- Data should be processed in place
- Orchestration of Data Ingestion
- Automated processes [2]

Model Quality Control: Information Flow and Activities Involved

Big data projects, including different organizations and many people, collecting and processing data require processes related to data quality control.

This quality control activities are implemented in the data storage in place (data lake/ ADL) and should be automated.

Some automatic activities are expected in ADL:

- Schedule control: the files have been uploaded according with a predefined schedule. If do not, some warnings should appear. Files not uploaded according with the schedule should also be pointed out.
- Completeness: no partial packages can be downloaded: all information regarding a model should be included (maps, reports, etc.). If do not, some warning should appear.
- Coding and naming: formal aspects regarding proper coding should be done automatically. If the information is not properly coded, some warnings should appear.
- General Dashboard, to be analyzed at a glance: uploaded information appears, with or without warnings.

This paper is about methodology and technical realization of big data environmental projects using automated workloads and quality control.

Methodology

Concept of the Structure

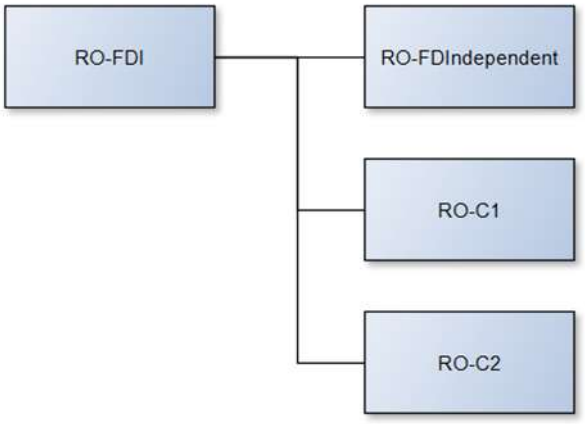
In this part of the study is explained the logical design of a prototypical data lake system for environmental resources, used for "Flood Hazard and Flood Risk Maps and Flood Risk Management Plans for Romania" project.

As solution for data management is proposed a structure of folders and files for storing the data collected and generated during FD Implementation Cycles. The proposed solution will be structured in such manner that will facilitate access to data generated in 1st Cycle of FD Implementation and help the user to identify the working folders and files that will be generated during 2nd FD Implementation Cycle. In this regard, the data will be organized at the first level of the structure based on the relationship of data with the Implementation Cycles of the Flood Directive in Romania. Based on various analysis the data will be organized in 3 main repositories as follows (Figure 2):

- Flood Directive Independent Data (RO-FDIndependent): data which were not generated in one of the projects related to Flood Directive implementation in Romania, but which can be used in the activities within the project frame as input data. These data are usually collected from different institutions, like INHGA, ANM, ANAR, ANCPI etc.
- Data under the FD umbrella:

- *RO-C1*: data collected and generated as deliverables in the Cycle 1 of FD, (survey data, hydrological data, models, maps, reports, etc.)
- *RO-C2* data generated as deliverables in the cycles of FD, in this case Cycle 2 FD (survey data generated during C2, hydrological data generated during C2, models, maps, reports, etc.)

Figure 2. *Data Structure at First Level*



For a better understanding of the data structure presented in this document, a mock-up of the entire data structure was developed and was populated with folders, sub-folders, and files. This mock-up also has the role to test the concept of data structure and naming convention before real data will be generated and stored according to it.

Data Organization Concept for Flood Directive Independent Data

The structure of container *RO-FDIndependent* is based upon grouping information by typology, UoM, rivers etc. Data in this container has following properties:

- generated by different institutions independent of the cycles of the FD Implementation.
- possibility to be used in more than 1 cycle of implementation.
- not oriented or grouped by the logic or necessities of any cycle of the FD; also, there will not be used concepts of the FD in grouping the data or in its content (e.g., APFSR, AFU or upstream-downstream to a certain unit of study, etc.).
- the refresh rate associated with this data differs from the logic of refresh of the institution that generates this data.
- the information should carry a time stamp and the refresh rate depends on the institution which administrates the data.

For a better understanding of data structure and content of FDIndepended container, in Figures 3 and 4 presents the general folder structure (tree) and types of content that could populate the vector container UoM level inside RO-FDIndependent.

At first level *FDIndependent* container folder is divided into 12 sub-folders, each one for a Unit of Management (UoM) plus one folder for National level. Besides those folders, another generic UoM (99-UoM) folder was created to be filled with sample data in the mock-up structure. This folder will not be present in the folder structure of the project.

Figure 3. *Data Structure at First Levels of the Container Folder RO-FDIndependent*

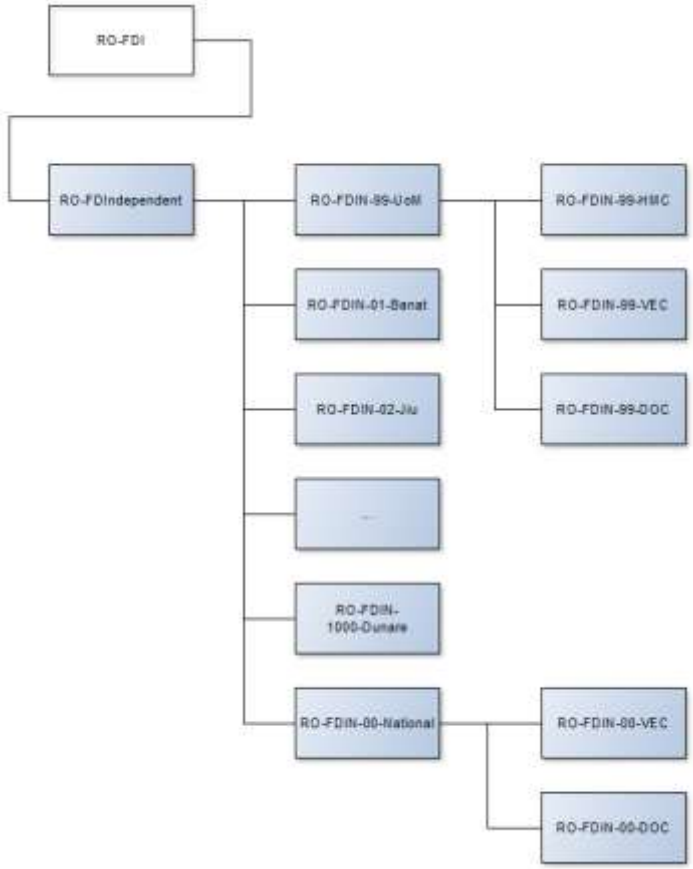
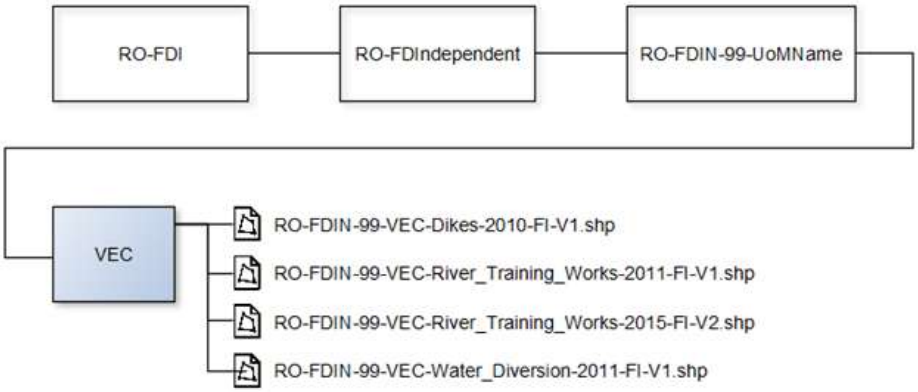


Figure 4. *Data Structure and Content at UoM Level of the Container Folder RO-FDIndependent*



In each UoM folder, there are three sub-folders containing the following three typologies of data: Hydro-Meteorological-Coastal data (HMC), GIS vector data (VEC) and documentations (DOC). The folder tree continues with different sub-types. The folder at National level contains only the last two type of folders and the specific of the information is at national level.

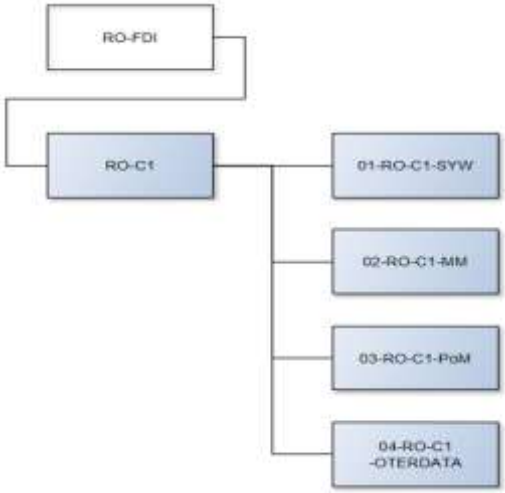
Data Organization Concept for Cycle 1

Data from Cycle 1 were organized based on the approach of every company contracted for executing the project. Therefore, the structure and the name of the deliverables do not follow a unique concept, not even at UoM level. Furthermore, for some RBAs the implementation of the Cycle 1 was done by two independent projects, by different companies at a few years distance between implementations, hence a lack of homogeneity of data structure, formats etc. Moreover, besides having a different structure and naming convention for each RBA, also the content differs significantly from one RBA to another, form input data, deliverables, intermediate data etc. Another issue was the presence of duplicates in the data received but located in different folders.

Cycle 1 was reorganized at a macro level, using a unique structure at a national level for each RBA. This process involved the following main types of actions:

- identification of the typology of data relevant to be used in Cycle 2 (e.g., survey data, models, etc.).
- reorganization of the relevant data, based on the newly developed unique structure concept.
- segregation of essential information from non-essential and removing duplicates.

Figure 5. *Data Structure at First Levels of the Container Folder RO-C1*



Data Organization Concept for Cycle 2

In comparison with Cycle 1, the current cycle of FD implementation will benefit of a data management system from the start. The structure and naming of products and deliverables will follow a unique concept developed at national level. The structure and future data in this repository have following properties:

- General structure data is largely based the main activities of the project (e.g., Modelling and mapping, PoM development, etc.).
- Data will be oriented or grouped by the logic or necessities required in 2nd Cycle of FD Implementation. The Cycle 2 data structure use concepts developed for Flood Directive for organizing the data (e.g., APSFR, AFU, etc.)
- the ability to dynamically group information based on their common properties reflected in the codes embedded in the name of folders and files.

The general structure for Cycle 2 data is largely based on the activities presented in the project Terms of Reference and can be regarded as a mapping of main products and deliverables that will be generated in current FD Implementation Cycle at National level.

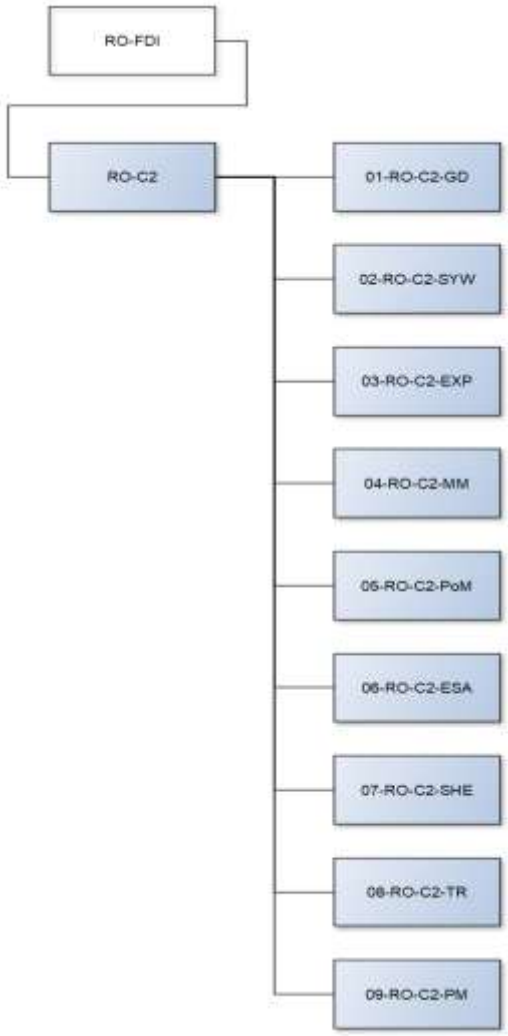
The sub-folders of RO-C2 folder indicate the main sections of the project as follows:

- 01-RO-C2-GD: General Documents (General documents such as ToR and annexes of the main contract, pilot studies, methodologies, etc.).
- 02-RO-C2-SYW: Survey Works (all survey works generated under Cycle 2 – DTM, Orto, DSM, Topological - Bathymetric data, etc.). Please note that, in this folder container, will be located only deliverables which are subject to survey works contracted during Cycle 2.
- 03-RO-C2-EXP: Exposure Data (geodatabase with all exposure data as shapefile).

- 04-RO-C2-MM: Modelling and Mapping (all relevant files and deliverables generated under this activity – models, shapefiles, raster's, pdf maps, reports, etc.).
- 05-RO-C2-PoM: Program of Measures (all deliverables and relevant files generated under this activity – screening reports, APSFR Strategies modelling and reports, IP modelling and reports, etc.)
- 06-RO-C2-ESA: Environmental and Social Aspects (SEA inputs);
- 07-RO-C2-SHE: Stakeholder Engagement (workshops related inputs, etc.).
- 08-RO-C2-TR: Trainings (training materials, data, etc.).
- 09-RO-C2-PM: Project Management (project management data - monthly reports, inception report, data management general description and user manual, etc.).

In Figure 6 is presented the firs level of data structure for Cycle 2 repository (RO-C2) based on the description:

Figure 6. *Data Structure at First Levels of the Container Folder RO-C2*



Based on analysis from 1st Cycle, the main bulk of products and deliverables in Cycle 2 will be generated by specific activities that will eventually populate the Survey Work container (02-RO-C2-SYW), Modelling and Mapping Container (04-RO-C2-MM) and Program of Measures container (05-RO-C2-PoM).

Data Management Approach

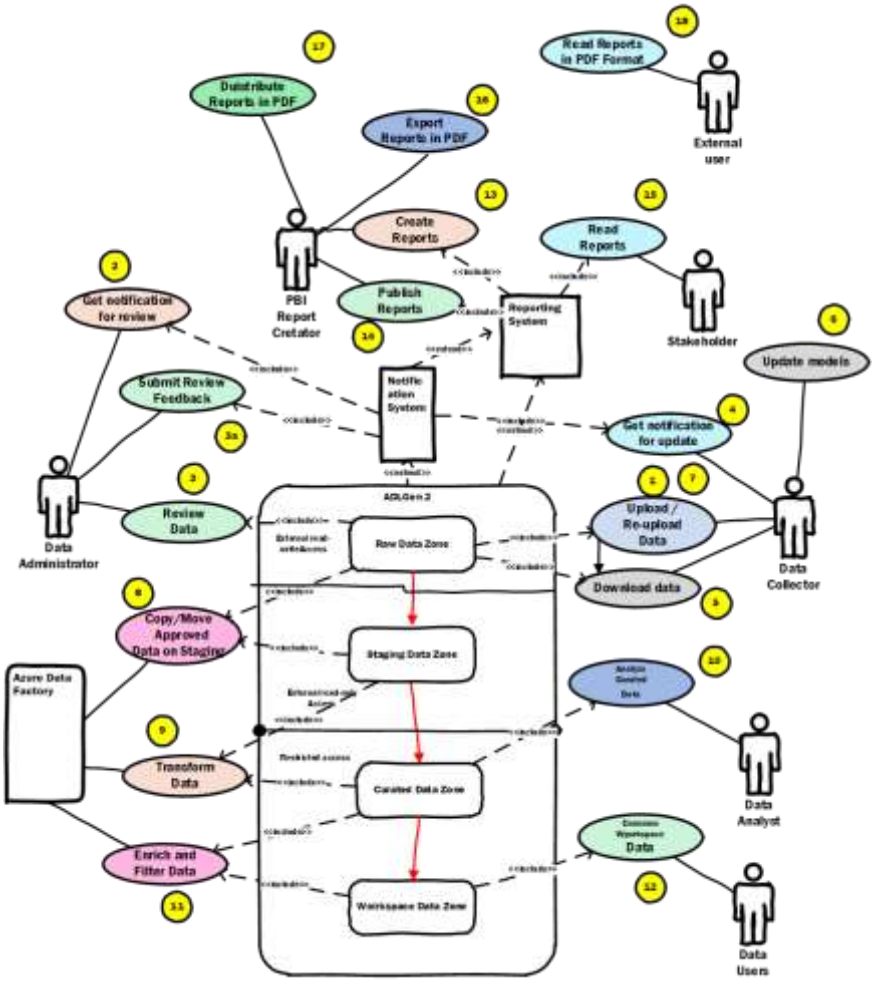
The current research has two parts:

- Common use-cases
- Overall process design flow technical implementation

Flood Hazard and Flood Risk Maps and Flood Risk Management Plans for Romania Use Cases

Common Use Cases

Figure 7. *ADL Common Use Cases*

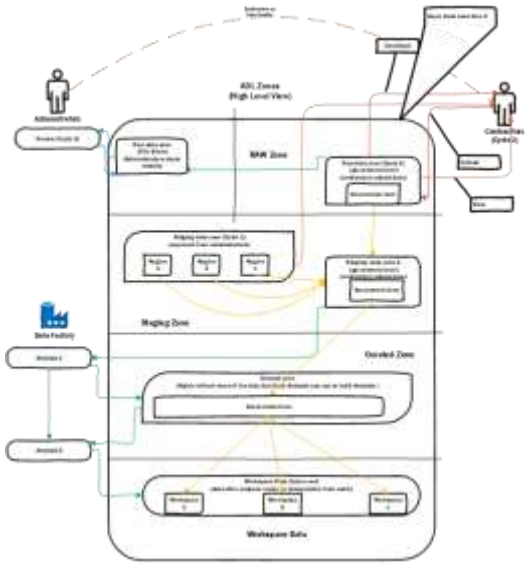


1. Data Collector Uploads Data into ADL (Raw Data Zone)
2. Data Administrator receive notification for new data upload (need analysis)
3. Data Administrator Review Uploaded data and fulfill forms for not approved models.
   a. Data Administrator Submit Review Feedback
4. Data collectors receive notification for not approved models.
5. Data collectors download models which need to be updated (Optional – they can have locally already created models and to not need to download it again).
6. Data collectors update models.
7. Data collectors re-upload models
8. Azure Data Factory moves approved data from Raw Data Zone to Staging Data Zone When?
9. Azure Data Factory transforms data structure from initial one by zones to a global one and write it into Curated Data Zone When?
10. Data Analyst creates analysis using data in Curated Data Zone
11. Azure Data Factory enriches and copy data into Workspace Data Zone (optional)
12. Data users consume data from Workspace Data Zone
13. Report Creator is creating a report based on Power BI and ADL data
14. Report Creator is publishing the report.
15. Stakeholder (with WBG account) is reading the published report
16. Report Creator is exporting reports in PDF format
17. Report creator distributes reports in PDF (email or push notification)
18. External user reads reports in PDF format

Overall Process Design Flow Technical Implementation

- Azure Data Lake zones:

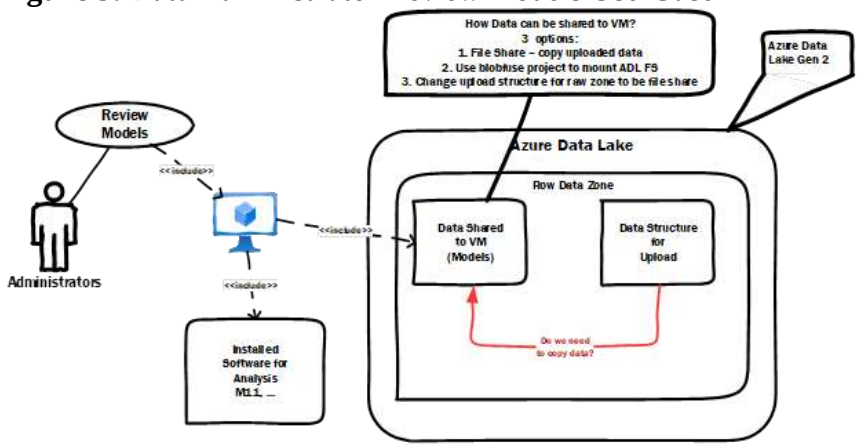
Figure 8. *Organization for Zones in ADL*



Technical Details for Use Cases

- Data Administrator Review Models
- Data Administrator Submit Feedback:
- Data collectors receive notification for not approved models.
- Data Collector Upload Models
- Data Administrator receive notification for uploaded models.

Figure 9. *Data Administrator Review Models Use-Case*



Notification Options

Notifications are very important for the effective workflows, related to data modeling and data management. There are several use-cases, technically implemented with flow automation and RPA,

Figure 10. *Notification Options*

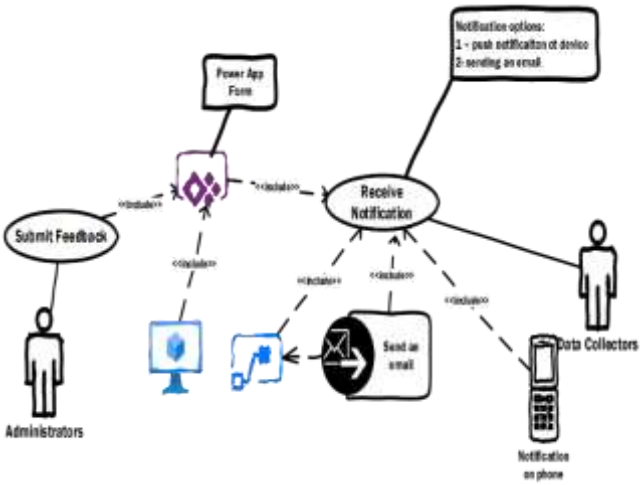
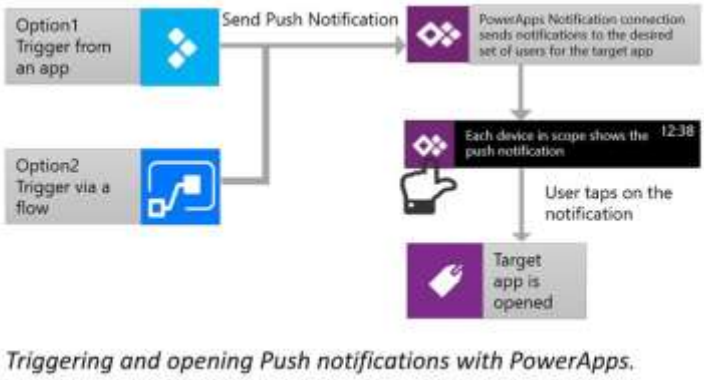


Figure 11. *Triggering Push Notifications with Power Apps*



Triggering and opening Push notifications with PowerApps.

Figure 12. *Data Administrator Receive Notification for Uploaded Models*

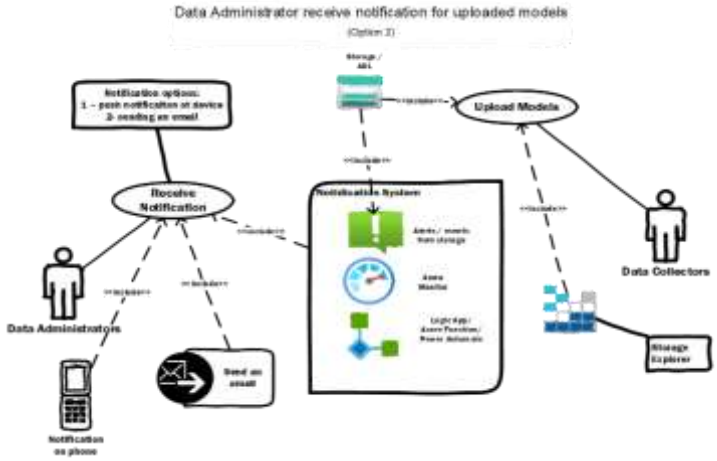
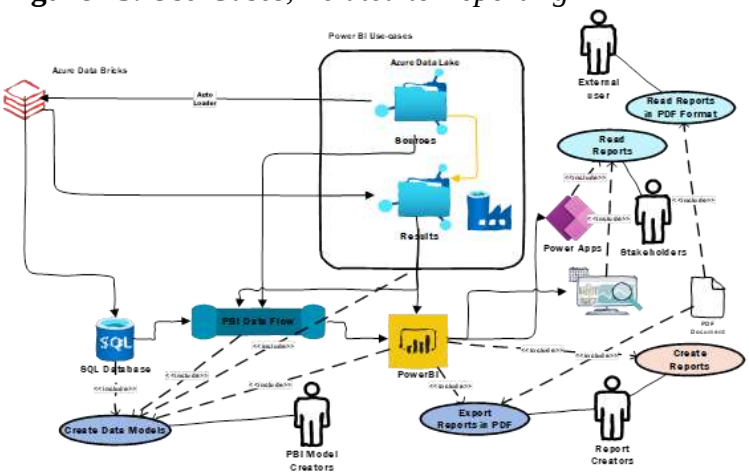


Figure 13. *Use-Cases, Related to Reporting*

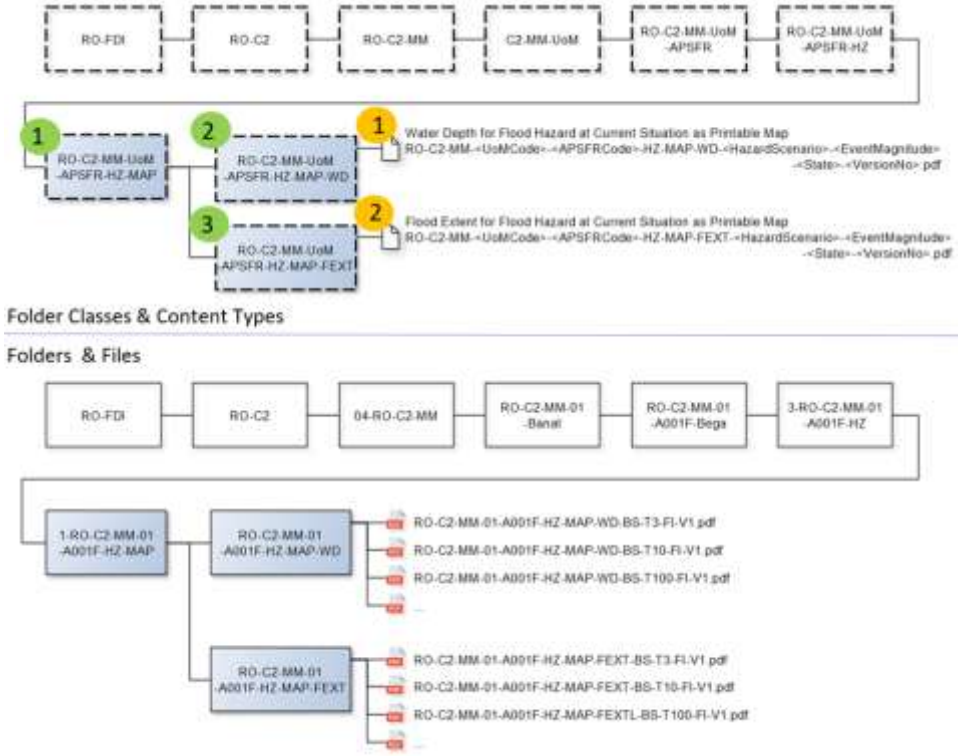


Data Structure Description and Naming Convention

The diagram presented in Figure 14 focuses on Content Type entity with examples for two distinct Content Types*. In relation with Content Types, the definition cards were also developed containing all predefined properties. The definition card for Content Type is similar with definition card for Folder Classes but with properties specific for files.

As a rule, the definition cards have a main role for defining the requirements for each deliverable imposing the naming convention for each content (as deliverable) but also stipulating quality specifications for each typology of deliverable. For a clear image on all the information a user would need to create files or folders that are compliant to the agreed standard and store them in the right location in the structure, in the diagram below is presented as example a section of the structure tree. The tree structure contains both Folder Classes and Content Types structure (top) as well as corresponding folders and file's structure

Figure 14. *Diagram with Examples of Folder Classes and Content Types (Upper Part) and the Corresponding Folders and Files (Lower Part)*



Results

In this paper, there are added new contributions about automation and optimization of data lake workloads, related to Flood Hazard and Flood Risk Maps and Flood Risk Management Plans for Romania

Table 1. *Project Parameters*

#	Data	Size
1	Total amount of data	200 TB +
2	Average data per zone	50 TB +
3	Package size	10 GB – 1 TB

Table 2. *Project Optimizations, Based on Automated Flows*

#	Operation	Improvement
1	Data Quality Assurance time	Decreased 3 times
2	Data upload time	Decreased 40%
3	Data publishing	Decreased 2 times
4	Project timeline optimization	Decreased 30%
5	Project cost optimization	Decreased 24%

Conclusions

The concept about automation of data lake flows in big data projects provides not only significant cost benefits but also minimize the time for delivery the projects, minimize risks and allow reusability of automation for future projects. Implementation of automated data lake flows make possible to extend scope for many projects allowing to process more data in reasonable timeline

Abbreviations

- AEP Annual Exceedance Probability
- ANAR The National Administration Romanian Waters
- ANM National Meteorological Administration
- APSFR Area of Potential Significant Flood Risk
- CT Content Type
- DEM Digital Elevation Model
- DTM Digital Terrain Model
- EAD Expected Annual Damage
- EC European Commission
- EU European Union
- FC Folder Class
- FD Flood Directive
- FRMP Flood Risk Management Plan
- GIS Geographical Information System

- RBA River Basin Authorities
- ToR Terms of Reference
- UoM Unit of Management

References

Data Lake Storage for Big Data Analytics | Microsoft Azure (2022). Available at: https://azure.microsoft.com/en-us/services/storage/data-lake-storage

Gupta, P., Ören, T. and Singh, M., 2019. Predictive intelligence using big data and the internet of things. IGI Global.

T. A. R. (auth.), Data Analytics: Models and Algorithms for Intelligent Data Analysis, 1st Edition, Vieweg+Teubner Verlag, 2012.

The Hitchhiker's Guide to the Data Lake (2022). Available at: https://azure.github.io/Storage/docs/analytics/hitchhikers-guide-to-the-datalake/

York, B. (2020) Trigger Automations from Azure Monitor Alerts - Cloud, Systems Management and Automation, Cloud, Systems Management and Automation. Available at: https://www.cloudsma.com/2020/02/trigger-automation-from-azure-monitor-alerts.

How to create a checklist in a Planner task with Power Automate (2021). Available at: https://tomriha.com/how-to-create-a-checklist-in-a-planner-task-with-power-automate.

Create and test an approval workflow with Power Automate. - Power Automate (2022). Available at: https://docs.microsoft.com/en-us/power-automate/modern-approvals.

Che, H. and Duan, Y. (2020) "On the Logical Design of a Prototypical Data Lake System for Biological Resources", Frontiers in Bioengineering and Biotechnology, 8. Doi: 10.3389/fbioe.2020.553904.

Llave, M. (2018) "Data lakes in business intelligence: reporting from the trenches", Procedia Computer Science, 138, pp. 516-524. Doi: 10.1016/j.procs.2018.10.071.