Cycle Orchestrator: A Knowledge-Based Approach for Structuring Cyclic ML Pipelines in the O&G Industry

Rafael Brandão[1], Vitor Lourenço[1], Marcelo Machado[1], Leonardo Azevedo[1], Marcelo Cardoso[1], Renan Souza[1], Guilherme Lima[1], Renato Cerqueira[1], Marcio Moreno[1]

[1] IBM Research, Rio de Janeiro, RJ, Brazil

Abstract. This work introduces the Cycle Orchestrator, a microservices infrastructure to structure and manage workflows related to heterogeneous data from the O&G domain. Through a knowledge-based perspective, it leverages reasoning, explainability and collaboration among stakeholders.

Keywords: Knowledge-based Workflow Orchestration, ML pipelines

Domain and requirements. In the natural resources domain, particularly in the oil and gas (O&G) industry, seismic data interpretation is key in exploration processes to identify geological structures in the subsurface, allowing experts to detect patterns and correlate geological factors by exploring different data sources. Commonly, this practice involves processing massive amounts of data through diverse techniques, aiming at detecting geological structures, enhancing information, correcting potential inconsistencies in the data acquisition process, and other purposes. An increasing number of works in the literature have been proposed applying Machine Learning (ML) workflows to support aspects of such processing. To systematically model geological exploration processes that apply complex data processing pipelines, allowing other stakeholders to collaborate and consume experiments' results, a holistic perspective is required. In this sense, we conceptualized and developed the Cycle Orchestrator, a knowledge-based workflow management system (WfMS) to support and operationalize the whole lifecycle of ML and general-purpose workflows. Including specification, setup, execution and provenance data management of such workflows. It was conceived within the O&G domain, primarily to support exploration use cases that apply cyclic ML workflows. Streams of tasks that can yield improved results through a chain of execution iterations. These workflows are associated to particular types of data sources (*e.g.* pre-stack and post-stack seismic data). The considered use cases comprised unsupervised ML pipelines that produce (train) new models and reuse pre-trained models and weights against new datasets, improving the quality by cyclic evolution. In this context, the orchestration involves the definition of what model and version should be applied to analyze a specific data source, as required by particular workflows applied in O&G exploration processes.

Knowledge-based Workflow Management. The Cycle Orchestrator takes advantage of the Hyperknowledge [2] conceptual model for relating knowledge specifications

aligned through a domain ontology to segments of multimodal content. Information is represented in the Hyperknowledge Base, a hybrid storage solution that uses a direct hyperlinked knowledge graph to maintain all information about workflow execution plans and provenance data stored in a knowledge base. The proposed modeling adheres to the MLWfM ontology [1] to structure basic aspects of ML and the PROV-ML [4] as provenance data model. Figure 1 shows the architectural overview of the system.

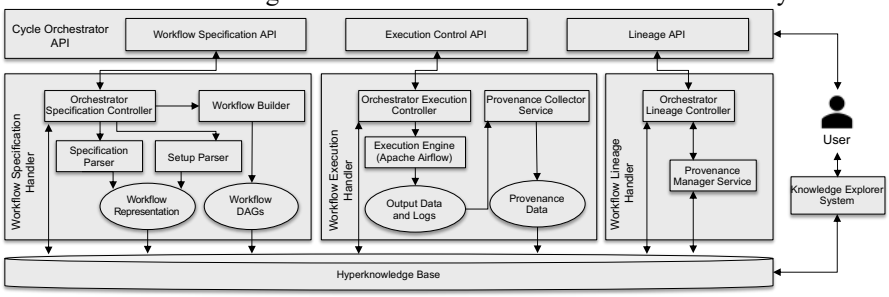

Fig. 1. Cycle Orchestrator's infrastructure overview.

Users interact with the system through a REST API and a web UI for curating and querying information, named Knowledge Explorer System (KES) [3]. The REST API has endpoints for workflow specification, execution, and lineage retrieval. The specification endpoints provide basic operations for workflow plans. Workflow definitions use a JSON-based specification language to model tasks, execution flow, required input data, expected output data and knowledge relations. This file is parsed, producing a Hyperknowledge representation and a directed acyclic graph (DAG) data structure. The Execution endpoint interfaces with the execution engine's API (Apache Airflow[1]). The execution handler captures provenance data, structuring according to the provenance data model that can be queried through the Lineage endpoint.

By integrating workflows' lifecycles in a common representation, our approach promotes knowledge production, consumption and curation in the O&G domain. Enabling industry experts to design exploration processes holistically, connecting heterogenous data processing, ontologies and stakeholders.

References

1. Moreno, M. et al.: Managing Machine Learning Workflow Components. In: 14th IEEE Conference on Semantic Computing, ICSC. pp. 25–30 (2020).
2. Moreno, M.F. et al.: Extending Hypermedia Conceptual Models to Support Hyperknowledge Specifications. Int. J. Semantic Computing. 11, 01, 43–64 (2017).
3. Moreno, M.F. et al.: KES: The Knowledge Explorer System. In: 2018 International Semantic Web Conference (P&D/Industry/BlueSky), ISWC. (2018).
4. Souza, R. et al.: Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering. In: 2019 IEEE/ACM Workflows in Support of Large-Scale Science, WORKS. pp. 1–10 (2019).

[1] https://airflow.apache.org/