Specification of Complex Analytics Workflows: A Formal Language Model of Decision Options

Pouriya Miri$^{1[0009-0008-8652-4855]}$, Petar Kochovski$^{1[0000-0003-4345-2069]}$, Marcela Tuler de Oliveira$^{2[0000-0001-5739-5590]}$, and Vlado Stankovski$^{1[0000-0001-9547-787X]}$

[1] University of Ljubljana, Faculty of Computer and Information Science, Slovenia
[2] Delft University of Technology, Department of Engineering Systems, Netherlands
pouriya.miri, petar.kochovski, vlado.stankovski@fri.uni-lj.si

Abstract. The specification of experiments expressed as Complex Analytics Workflows is a complex task that involves many decision-making steps with various degrees of complexity. The use of the context, the expert knowledge, and the potential for its sharing and reuse in the context of experiment specification have not been addressed sufficiently until now. Moreover, to make such knowledge instrumental, it should be coupled with specific probabilistic measures, such as particular assurances, ranking, and verification of various options. The paper aims to present a novel semantic model for probabilistic reasoning in any experimentation context coupled with a functional system for knowledge generation, reuse, and sharing. The result of our work can be used within existing experimentation engines.

Keywords: Semantic model· Complex Analytic Workflow· · Markov Decision Process.

1 Introduction

Complex Analytics Workflows (CAWs) provide an advanced framework for managing data-driven analytics, leveraging technologies such as Artificial Intelligence (AI) and Machine Learning (ML). These workflows are crucial for handling multi-faceted tasks that require precision, flexibility, and integration of heterogeneous data sources. Unlike traditional workflows, CAWs are designed to adapt and optimize scientific experiments, simulations, and real-world scenario analyses, incorporating feedback and learning mechanisms throughout the process.

The importance of CAWs is highlighted in projects like ExtremeXP[1], demonstrating their ability to fit AI/ML models to specific tasks, involving humans in the control loop to achieve the highest possible accuracy and precision. However, specifying CAWs requires expert knowledge, particularly when selecting data and defining experiment configurations.

This paper proposes a semantic model for CAWs that incorporates probabilistic reasoning through a Markov Decision Process (MDP), facilitating knowledge

[1] https://extremexp.eu/

generation, sharing, and reuse. The model is designed to be integrated into existing workflow tools, like Taverna[2], or Ascalon[3], enhancing their capability to manage complex decision-making processes with formal assurances, ranking, and verification. In Section 2, we provide background on CAWs. Section 3 analyzes a public administration use case that could be fit on the high-level CAW. Section 4 elaborates on the semantic language model that represents the concepts that are involved in the EMF model for the CAW. Finally, Section 5 discusses the current development in the context of the aims of the ExtremeXP project.

2 Background

The evolution of workflows, particularly in handling complex, multi-step analytic tasks, has been a focal point of research since Casati et al. (1970) [1] introduced the conceptual modeling of workflows. Gil et al. (2013) [2] advanced this field by exploring the dynamic configuration of workflows to manage large datasets, which has become increasingly relevant in today's data-driven environments. Deelman et al. (2016) [3] contributed to this evolution by introducing performance modeling and diagnostic approaches for extreme-scale workflows, emphasizing the importance of assessing workflow efficiency.

In modern technological advancements, the Horizon Europe ExtremeXP project redefines workflows within the framework of IoT, Blockchain, AI, Cloud-to-Edge computing, and Digital Twins. These technologies are pivotal in developing and executing CAWs, essential for achieving high precision in scientific experiments and decision-making processes. Krishnan et al. (2021) [4] and Forkan et al. (2023) [5] have also contributed by developing workflows tailored to specific applications, such as germline variant calling and spatial analytics, respectively, highlighting the diverse applicability of CAWs. Oliveira et al. (2022) [7] proposed a transparent access control mechanism using Attribute-Based Access Control (ABAC) integrated with blockchain technology. This model offers a sophisticated approach to security, providing fine-grained access control, accountability, and transparency in CAWs. Their work is instrumental in ensuring that workflows can be securely shared, reused, and monetized, making it a cornerstone of decentralized knowledge systems.

Integrating a semantic model of CAWs with an MDP, as pursued in the ExtremeXP project, builds upon these advancements. By capturing, sharing, and reusing expert knowledge in experiment specification, this integration ensures that workflows are precise, adaptable, and capable of providing formal assurances, ranking, and verifiable results.

3 High-Level CAW Knowledge System Design

This section analyses complex interactions among expert users and a workflow system. An example of a CAW is a workflow in which the user intends to re-

[2] http://www.taverna.org.uk/

[3] https://ascalon.fr/

alize a Sustainable Development Goal (SDG)[4]. Workflows of this kind can be implemented in many domains, such as science, engineering, and public administration. In a CAW, complex decision-making involves evaluating various options, considering uncertainties and probabilities, making optimal choices in the short term, and aligning with long-term sustainability goals. In this context, sustainable use cases involve decision options that contribute to the UN's SDGs, minimize negative environmental impacts, promote social responsibility, and support long-term economic viability.

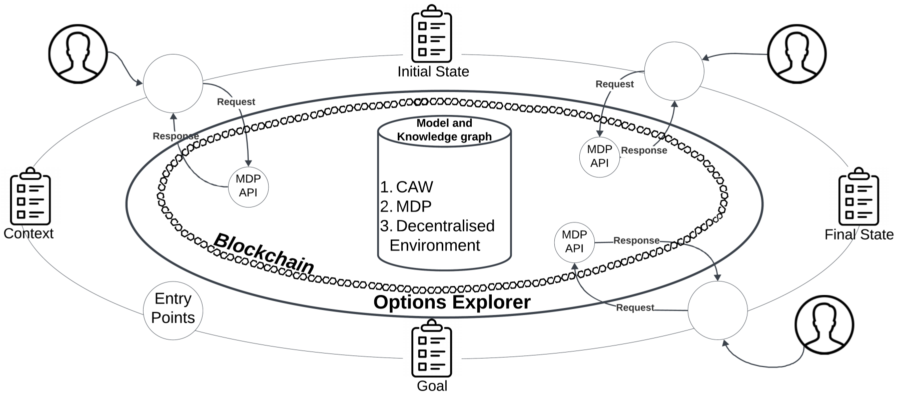

Fig. 1. A decentralized ecosystem for expert users of Complex Analytic Workflows

As illustrated in Figure 1, a semantic model of (1) CAWs, (2) MDP, and (3) Smart Contracts allows expert users to generate, share, reuse, and otherwise exchange knowledge about workflows in a decentralized manner. Our decentralized system assists expert users in navigating through various states in different use cases by accessing the knowledge stored in a knowledge graph. The CAW employs MDP method to suggest ranked sequences based on users' entry points and desired outcomes. With this workflow, users can obtain answers to their needs with diverse levels of knowledge and varying degrees of specificity within the system. Users who begin with specific initial states are guided through sequences of actions that lead to potential outcomes. The system analyzes the current state, identifies available transitions, and suggests sequences that maximize the likelihood of achieving optimal outcomes. The machine identifies potential starting points and suggests sequences of actions that lead to the known final state. Users who know their goals but lack knowledge about specific starting or ending points are guided through sequences that bridge the gap between their current knowledge and desired objectives. The decision-maker assesses the user's goal, identifies relevant starting points and potential outcomes, and suggests actions

[4] https://sdgs.un.org/goals

aligning with the user's objectives. In the following, we present a use case of smart public administration that motivated our research.

Each government aims to address specific citizen issues. Currently, integrating multiple public services into effective workflows is inefficient. Successful processes should be recorded, stored in a knowledge base, and used to resolve new cases based on similarity. By doing this, public administration decision-makers can create efficient workflows that leverage various services to address citizen situations. They can record successful processes, store them in a knowledge base, and access them to resolve new cases. This enhances efficiency by making effective methods readily available and adaptable to each citizen's unique needs. The model helps decision-makers assess applicant eligibility and compliance, prioritize evaluations, and expedite processes. Applicants can select workflows, prepare requirements, and verify completeness. Verification protocols ensure data integrity and security, building trust. Officials can analyze historical data to identify effective processes, speed up query resolution, and optimize resources.

4 Semantic Model for Probabilistic Reasoning

The Eclipse Modeling Framework (EMF) offers a robust infrastructure for developing structured data models, defining model structures, maintaining instances, and generating code. Our model leverages EMF as a foundational tool to orchestrate complex analytical processes, addressing the challenges organizations face with growing data sources and increasingly intricate tasks.

The CAW EMF model provides a comprehensive framework for designing and executing complex analytical workflows. It incorporates six fundamental concepts that form the basis for constructing detailed analytical models, guiding users systematically from data intake through to insight generation.

Table 1. The concept of complex analytics workflow and the concept of MDP state.

Concept: Complex Analytics Workflow			**Concept: MDPState**		
Properties	**Range**	**Cardinality**	**Properties**	**Range**	**Cardinality**
:hasMDPState	:MDPState	1...*	:hasContextFeature	:ContextFeature	1...*
:hasContextFeature	:ContextFeature	1...*	:hasInitialState	xsd:boolean	1
:hasTransition	:Transition	1...*	:hasFinalState	xsd:boolean	1
:isSequential	xsd:boolean	1	:hasGoal	xsd:boolean	1
:isConsensual	xsd:boolean	1	:hasReward	xsd:boolean	1
:hasExpertName	xsd:string	1	:hasIndicators	xsd:string	1

The CAW concept, detailed in Table 1, encompasses all possible MDP states, features, and transitions within a workflow. CAWs may be performed by one or more experts, adhering to a sequence of transitions where the property isSequential is set to true. In cases without a specific sequence, isSequential is set to false. When user choices are involved, the property isConsensual is set to true. The MDPState concept, also outlined in Table 1, represents the current state within an MDP model. Each MDP state can exhibit characteristics like being an initial state, final state, goal, or offering a reward, with relevant properties set to true. MDP states may hold multiple actual properties simultaneously. As shown in Table 2, indicators quantify aspects such as accuracy, precision, and user scores, providing insight into state dynamics. These indicators, expressed as integer values, reflect varying performance and engagement levels. MDP states are defined by features that describe the state's attributes and dynamics. The transition concept defines the movement from one state to another within a CAW. These transitions, integral to the CAW framework, align with predefined workflow sequences and are crucial in guiding system dynamics and decision-making processes. Transitions also incorporate MDP actions, which are tailored to respond to specific trends, whether increasing or decreasing, ensuring that the workflow adapts to changing conditions.

Table 2. The concepts of indicators, MDP action, and transition.

Concept: Indicators		**Concept:Transition**		**Concept:MDPAction**	
Properties	**Range**	**Properties**	**Range**	**Properties**	**Range**
:hasAccuracy	xsd:int	:hasStartState	xsd:string	:hasAccuracyTrend	xsd:string
:hasPrecision	xsd:int	:hasEndState	xsd:string	:hasPrecisionTrend	xsd:string
:hasUserScore	xsd:int	:hasMDPAction	xsd:string	:hasScoreTrend	xsd:string

The operational workflow of the CAW system highlights the interactions among various components involved in decision-making and knowledge management. Domain experts with digital credentials (1) contribute by adding states to the knowledge graph, enriching the system with their expertise. AI algorithms (2) continuously monitor the system's performance, recommending updates to optimize efficiency. Entity Embeddings (3) store all available states, providing necessary information when requested by the MDP machine (4), which acts as the system's decision-making core. The MDP machine simulates and evaluates actions, interacting with the reward simulator (5) to validate outcomes. Users (6) engage with the system by sending requests to the server (7), facilitating interaction and monitoring CAW performance. The MDP machine generates and ranks user action recommendations, ensuring informed decision-making.

5 Discussion and Conclusions

This paper introduces a new Options Explorer that is designed to aid expert users in exploring various solutions to a given problem. Experts' decisions are influenced by their goals, intentions, context, and method limitations, often made under complex, non-deterministic conditions. The paper emphasizes the importance of the Options Explorer, highlighting possible decision options and providing probabilistic assurances, ranking, and verification based on previous executions. The model comprises three sub-models: (1) a CAW, (2) an MDP model, and (3) Blockchain-based Smart Contracts environment. These components enable decentralized knowledge generation, sharing, and reuse among expert users stored in a knowledge graph. The model offers assurances, ranking, and verification of options and supports trustworthiness, transparency, traceability, and access control. This ongoing work under the ExtremeXP research project aims to develop a prototype for integrating existing workflow systems.

Acknowledgment

The research and development reported in this paper have received funding from the European Union's Horizon Program under grant agreement no. 101093164 (ExtremeXP: EXPerimentation driven and user eXPerience oriented analytics for eXtremely Precise outcomes and decisions).

References

1. Casati, F., Ceri, S., Pozzi, Giuseppe. (1970). Conceptual Modeling of WorkFlows. https://doi.org/10.1007/BFb0020545.
2. Yolanda Gil, Varun Ratnakar, Rishi Verma, Andrew Hart, Paul Ramirez, Chris Mattmann, Arni Sumarlidason, and Samuel L. Park. 2013. Time-bound analytic tasks on large datasets through dynamic configuration of workflows. In Proceedings of the 8th Workshop on Workflows in Support of Large-Scale Science (WORKS '13). Association for Computing Machinery, New York, NY, USA, 88–97. https://doi.org/10.1145/2534248.2534257
3. Deelman E, Carothers C, Mandal A, et al. PANORAMA: An approach to performance modeling and diagnosis of extreme-scale workflows. The International Journal of High Performance Computing Applications. 2017;31(1):4-18. https://doi.org/10.1177/1094342015594515
4. Krishnan, V., Utiramerur, S., Ng, Z. et al. Benchmarking workflows to assess performance and suitability of germline variant calling pipelines in clinical diagnostic assays. BMC Bioinformatics 22, 85 (2021). https://doi.org/10.1186/s12859-020-03934-3
5. Forkan, A., Both, A., Bellman, C., Duckham, M., Anderson, H., and Radosevic, N. (2023) K-span: Open and reproducible spatial analytics using scientific workflows. Frontiers in Earth Science. https://doi.org/10.3389/feart.2023.1130262
6. Merkle, N., Mikut, R. (2023) Context-Aware Composition of Agent Policies by Markov Decision Process Entity Embeddings and Agent Ensembles. arXiv eprint=2308.14521. https://doi.org/10.48550/arXiv.2308.14521
7. M. Tuler De Oliveira, L. H. A. Reis, Y. Verginadis, D. M. F. Mattos and S. D. Olabarriaga, "SmartAccess: Attribute-Based Access Control System for Medical Records Based on Smart Contracts," in IEEE Access, vol. 10, pp. 117836-117854, 2022, doi: 10.1109/ACCESS.2022.3217201.