Semantic Cloud System for Scaling Data Science Solutions for Welding at Bosch

Zhuoxun Zheng[1,2], Baifan Zhou[3,2], Zhipeng Tan[1,4], Ognjen Savkovic[6], Diego Rincon-Yanez[1,7], Nikolay Nikolov[5,2], Dumitru Roman[5,3], Ahmet Soylu[3,2] and Evgeny Kharlamov[1,2]

[1]*Bosch Center for AI, Germany*
[2]*Department of Informatics, University of Oslo, Norway*
[3]*Department of Computer Science, Oslo Metropolitan University, Norway*
[4]*RWTH Aachen University, Germany*
[5]*SINTEF AS, Norway*
[6]*Free University of Bozen-Bolzano, Italy*
[7]*Universidad de Santander, Cucuta, Colombia*

Background and Challenges. Industry 4.0 focuses on smart factories that rely on IoT technology for automation. This produces massive amounts of production data, increasing the demand for data-driven solutions and cloud technology. Yet, users of these solutions and cloud technology are often not cloud experts, such as domain experts and data scientists (Fig. 1.1). In a standard setting of a data science project, the team requires extensive assistance from cloud experts, whenever they want to deploy solutions or make small changes to their solutions deployed on the cloud. To facilitate the adoption of cloud systems careful planing to balance cost and benefits is required. Scaling data science solutions presents challenges of handling high data volume and enabling a broader users which are non-cloud experts to use cloud systems.

SemCloud for Distributed ETL and Use Case. In industry, large amount of data collected from different resources are integrated and analysed in parallel to optimise the following production. Due to the large volume of data, cloud technology is used to enable distributed ETL. Here, cloud configuration plays an important role in achieving optimal performance, which is however non-trivial for non-cloud experts. To address the scalability challenges and democratising cloud systems for more users, we propose SemCloud [1], a semantics-enhanced cloud system, that scales semantic ETL pipeline on the cloud, and allows non-cloud experts to deploy their solutions. We showcase SemCloud in our welding use case.

SemCloud for Automated Cloud Configuration. SemCloud achieves optimised cloud adoption [2] for ETL automatically by breaking down the ETL into pipelines of four steps: *retrieve*, *slice*, *prepare*, and *store* (Fig. 1.2), where data is first retrieved from databases or online streams, and then split into subsets (e.g. each belong to one welding machine) by *slice* to achieve parallel processing and storage in the following *prepare* and *store*. A rough description of the application of SemCloud is as follows: (a) non-cloud experts create knowledge graphs (KG) that represent ETL-Pipelines on a cloud system, where attributes of cloud resource configuration is under-specified; (b) Datalog ML rules execute in three steps, where the rules contain external

ISWC2023: The 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece

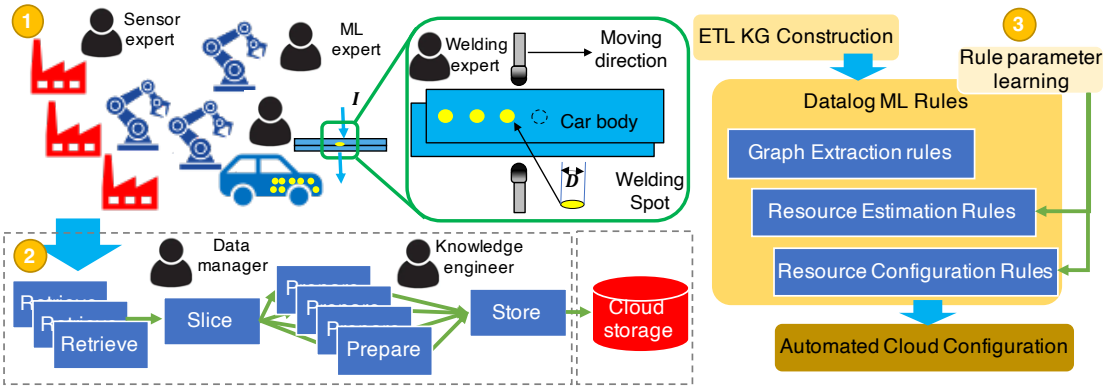

Figure 1: (1) High volume of heterogeneous data collected from welding machines and factories; (2) distributed semantic ETL; (3) ETL KG and Datalog ML rules for automated cloud configuration

functions obtained by (c) rule parameter learning with ML; (d) automated cloud configuration.

SemCloud Ontology and KG for ETL Pipelines. SemCloud provides the users an ontology to construct semantic ETL pipelines and encode them into knowledge graphs. The ontology is written in OWL 2, and consists of 20 classes and 165 axioms. For these data, the users construct KG for ETL pipelines with four layers (via GUI), which will be used for rule-based reasoning.

Datalog ML Rules. Obtaining an optimised cloud configuration is not trivial. Cloud experts typically try different configurations by testing the system with various settings and use heuristics to manually decide on the configurations. To this end, SemCloud uses adaptive rules in *Datalog* with aggregation and calls to external predicates learned by ML . In particular, we consider non-recursive rules of the form $B \leftarrow B_1, \ldots, B_n$, where B is a head of rule (the consequence of the rule application) and $B_1, \ldots, B_n$ are either predicates that apply join, aggregate function that filters out the results or the expression of the form $Var = @FUNCT(Vars)$.

Rule Parameter Learning with ML. The functions in the adaptive rules are in the form of ML models. The *resource estimation rules* are selected from the best model resulting from training three ML methods and the pilot running statistics. We selected three representative classic ML methods: *Polynomial Regression (PolyR)*, *Multilayer Percetron (MLP)*, and *K-Nearest Neighbours (KNN)*. The *resource configuration rules* are trained with the three ML methods and with optimisation techniques, such as Bayesian optimisation or grid search.

User Feedback and Business Impact. SemCloud helps non-cloud experts who know little about cloud and cannot use cloud system to find the optimal allocation of cloud resources in various industrial tasks. To verify the time efficiency of SemCloud, we run SemCloud repeatedly 3562 times and gather pilot running statistics. With SemCloud, the Bosch semantic ETL is speed up to at least twice faster, the optimisation time of cloud configuration is speeded up to 1.12s. Additionally SemCloud helps more users to use cloud systems, which greatly reduce time and cost for personnel training and data processing, benefiting data science solution at Bosch.

References

[1] B. Zhou, et al., Scaling data science solutions with semantics and ML, in: ISWC, 2023.
[2] Z. Zheng, O. Savkovic, et al., Datalog with external machine learning functions for automated cloud resource configuration, in: ISWC, 2023.