Contribution ID: **148** Type: **Paper presentation**

Semantic-Aware Data Lineage Tracking for Transparent ETL Pipelines in Industrial Systems

Thursday 18 September 2025 11:40 (15 minutes)

In modern industrial environments, understanding how data flows through complex ETL pipelines is critical for traceability, auditing, and compliance. While traditional lineage tracking tools rely on static metadata or log-based introspection, they often lack semantic expressiveness and offer limited support for automated validation.

This work presents a semantic-aware framework for ETL data lineage tracking, integrating property graph modeling (Neo4j), lightweight ontologies (RDF/OWL), and SPARQL querying. Each transformation step is modeled as a semantic entity enriched with metadata such as inputs, outputs, timestamps, and execution order. The resulting lineage graph is exported to RDF, enabling validation via SHACL and pattern-based queries over transformation chains.

The proposed solution was implemented and tested on the AdventureWorks DW2022 dataset, demonstrating low-latency querying, structural correctness, and enhanced traceability. Comparative analysis shows that, unlike traditional tools such as Apache Atlas or OpenLineage, our approach supports fine-grained reasoning, constraint checking, and semantic completeness verification.

This contribution offers a lightweight and reproducible solution using only open-source components and is particularly suited for evolving, compliance-heavy industrial data environments. Future work includes extending the ontology, integrating streaming ETL sources, and deploying the model over scalable RDF triple stores.

Author: BINDEA, Bogdan Nicușor (Technical University of Cluj-Napoca, Computer Science Department)

Co-authors: CENAN, Călin (Technical University of Cluj-Napoca, Computer Science Department); DÎNȘORE-ANU, Mihaela (Technical University of Cluj-Napoca, Computer Science Department)

Presenter: BINDEA, Bogdan Nicușor (Technical University of Cluj-Napoca, Computer Science Department)

Session Classification: High Performance Computing in Science

Track Classification: Cloud Computing and Network Virtualisation