Toward Executing Datalog on Big Data Platforms - Extended Abstract

Andrea Cuteri, Giuseppe Mazzotta and Francesco Ricca

Department of Mathematics and Computer Science, University of Calabria, Rende, Italy

Nowadays database systems face growing demands to handle enormous datasets produced by a wide range of applications. Apache Spark has become a popular framework for large-scale data analytics, providing effective in-memory computation capabilities over distributed systems. When developing analytics solutions in Spark, developers must use imperative programming approaches to express the execution logics of queries. In contrast, declarative query languages provide more abstract interfaces that enable users to define desired outcomes without specifying the computational procedures. Datalog has established itself as a fundamental declarative language in database systems, supporting the expression of sophisticated queries through a logic-based syntax. Nevertheless, conventional Datalog implementations cannot handle Big Data workloads effectively because they lack the capability to exploit distributed computing architectures. Efforts to address these limitations have led to the creation of next-generation Datalog engines built upon scalable frameworks such as Spark. We introduce *Datalog2Spark*, a prototype system that investigates compilation of Datalog programs into Spark applications for execution on computing clusters. In our prototype implementation, we concentrate on non-recursive Datalog programs featuring stratified negation, enhanced with data type support and several beneficial aggregation operations. These enhancements were developed recognizing that (i) data type support is essential in database applications, and (ii) aggregation capabilities have proven vital for successful data processing and analysis tasks.

Therefore, we expand the Datalog language by incorporating type definition constructs as well as aggregation functions, such as average, median, etc., specifically designed to meet practical use-cases in Big Data analytics. To evaluate *Datalog2Spark*'s performance, we conducted an experimental evaluation in which we compared *Datalog2Spark*against two in-memory engines for Datalog : DLV and CLINGO.Obtained results confirm the viability of the approach, demonstrating its potential to bring the benefits of declarative programming to Big Data analytics.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

ASPOCP 2025: 18th Workshop on Answer Set Programming and Other Computing Paradigms September 12-13th, 2025
andrea.cuteri@unical.it (A. Cuteri); giuseppe.mazzotta@unical.it (G. Mazzotta); francesco.ricca@unical.it (F. Ricca)
0009-0000-7629-7347 (A. Cuteri); 0000-0003-0125-0477 (G. Mazzotta); 0000-0001-8218-3178 (F. Ricca)