Enhancing VLAM Workflow Model with MapReduce Operations

Mikołaj Baranowski\textsuperscript{a}, Adam Belloum\textsuperscript{a}, Marian Bubak\textsuperscript{a,b}

\textsuperscript{a}Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
\textsuperscript{b}Department of Computer Science, AGH University of Science and Technology, Mickiewicza 30, 30-059 Kraków, Poland

{baranowski, a.s.z.belloum}@uva.nl, bubak@agh.edu.pl

Motivation

- Importance of MapReduce in processing big data
- Pig Latin and Sawzall – solutions based on Domain Specific Languages that provide simple and user-friendly access to MapReduce resources
- To get access to MapReduce resources, users have to use different environments for specifying and running MapReduce jobs along with other application models like workflows

Design and Implementation

- Designed DSL describes only Map operation
  - Map operation is changed many times during the implementation process and the most of the execution time is spend on waiting for I/O operations
  - Users rarely change reduce and aggregate operations and they use a small number of them
  - The execution time strongly depends on reduce phase
- DSL translates Map operations to many platforms
  - Specifies types of processed data (required statically typed Hadoop reducers)
  - Defined with Ruby programming language which allows to choose an appropriate implementation

Example application (word count)

- DSL was used to define Map operation (Listing 1)
- Special routines (map, c.string, c.number) were designed to simplify development of Map operations
- Sum reducer (c.sum) is included in the Hadoop distribution

Conclusions

- Comparing to others, developed method provides a portable and pluggable solution
- The solution based on dynamic languages and DLS allows to define Map operation with a short, clear code
- It can be adapted to many existing applications thanks to the limited number of dependencies (Ruby)
- Map operations defined with the proposed DSL can be executed on many MapReduce platforms

Acknowledgements – This work was supported by the Dutch National Program COMMIT.

Bibliography