# Scalable Script-based Data Analysis Workflows on Clouds

Fabrizio Marozzo
DIMES
University of Calabria
Italy
fmarozzo@dimes.unical.it

Domenico Talia
DIMES
University of Calabria
ICAR-CNR
Italy
talia@dimes.unical.it

Paolo Trunfio
DIMES
University of Calabria
Italy
trunfio@dimes.unical.it

## ABSTRACT

Data analysis workflows are often composed by many concurrent and compute-intensive tasks that can be efficiently executed only on scalable computing infrastructures, such as HPC systems, Grids and Cloud platforms. The use of Cloud services for the scalable execution of data analysis workflows is the key feature of the Data Mining Cloud Framework (DMCF), which provides a Web interface to develop data analysis applications using a visual workflow formalism. In this paper we describe how we extended DMCF to support also the design and execution of script-based data analysis workflows on Clouds. We introduce a workflow language, named JS4Cloud, that extends JavaScript to support the implementation of Cloud-based data analysis tasks and the handling of data on the Cloud. We also describe how data analysis workflows programmed through JS4Cloud are processed by DMCF to make parallelism explicit and to enable their scalable execution on Clouds. Finally, we present a data analysis application developed with JS4Cloud, and the performance results obtained executing the application with DMCF on the Windows Azure platform.

## Categories and Subject Descriptors

Computer systems organization [**Architectures**]: Distributed architectures—*Cloud computing*

## Keywords

Workflows, Data analysis, Data mining, Cloud computing, Scalability, JS4Cloud

## 1. INTRODUCTION

Workflows are an effective paradigm to address the complexity of scientific and business applications. They provide a declarative way of specifying the high-level logic of an application while hiding the low-level details that are not fundamental for application design [14][13]. The use of workflows has proven to be very effective to describe complex data analysis processes, e.g. Knowledge Discovery in Databases (KDD) applications, which can be conveniently modelled as graphs linking together data sources, filtering tools, data mining algorithms, and knowledge models.

Data analysis workflows are often composed by many concurrent and compute-intensive tasks that can be efficiently handled only on scalable computing infrastructures, such as HPC systems, Grids and Cloud platforms. The use of Cloud services for the scalable execution of data analysis workflows is the key feature of the *Data Mining Cloud Framework* (DMCF), presented in [9]. In DMCF, data analysis workflows are designed through visual programming, which is a very effective design approach for high-level users, e.g. domain-expert analysts having a limited understanding of programming. In addition, a graphical representation of workflows intrinsically captures parallelism at the task level, without the need to make parallelism explicit through control structures [7]. On the other hand, script-based workflows can be used as an effective alternative to graphical workflows, since the formers can allow expert users to program complex applications more rapidly, in a more concise way, and with higher flexibility. Therefore, we extended the DMCF system to support also script-based data analysis workflows, as an additional and more flexible programming interface for skilled users.

In this paper we describe our solution for programming and executing parallel script-based data analysis workflows in DMCF. We introduce a workflow language, named JS4Cloud, that extends JavaScript to support the development of Cloud-based data analysis tasks and the access to data on the Cloud. The main benefits of JS4Cloud are: *i*) it is based on a well known scripting language, so that users do not have to learn a new programming language from scratch; *ii*) it implements a data-driven task parallelism that automatically spawns ready-to-run tasks to the available Cloud resources; *iii*) it exploits implicit parallelism so application workflows can be programmed in a totally sequential way, which frees users from duties like work partitioning, synchronization and communication.

The remainder of the paper is organized as follows. Section 2 shortly presents the Data Mining Cloud Framework and its visual workflow formalism. Section 3 presents JS4Cloud and discusses how workflows programmed through this language are executed by DMCF. Section 4 describes a data analysis application developed with JS4Cloud, and presents performance results obtained executing the application with DMCF on the Windows Azure platform. Sec-