TITAN: A knowledge-based platform for Big Data workflow management

Antonio Benítez-Hidalgo[1,2], Cristóbal Barba-González[1,2], José García-Nieto[1,2], Pedro Gutiérrez-Moncayo[1,2], Manuel Paneque[1,2], Antonio J. Nebro[1,2], María del Mar Roldán-García[1,2], José F. Aldana-Montes[1,2], and Ismael Navas-Delgado[1,2]

[1] Grupo de investigación Khaos, Dpto. Lenguajes y Ciencias de la Computación
`antonio.b@uma.es`, {`cbarba,jnieto`}`@lcc.uma.es`
`mpaneque@uma.es`, {`antonio, mmar,jfam,ismael`}`@lcc.uma.es`
[2] ITIS Software, Universidad de Málaga, Málaga, España.

Resumen(Abstract). Modern applications of Big Data are transcending from being scalable solutions of data processing and analysis, to now provide advanced functionalities with the ability to exploit and understand the underpinning knowledge. This change is promoting the development of tools in the intersection of data processing, data analysis, knowledge extraction and management. In this paper, we propose TITAN, a software platform for managing all the life cycle of science workflows from deployment to execution in the context of Big Data applications. This platform is characterised by a design and operation mode driven by semantics at different levels: data sources, problem domain and workflow components. The proposed platform is developed upon an ontological framework of meta-data consistently managing processes and models and taking advantage of domain knowledge. TITAN comprises a well-grounded stack of Big Data technologies including Apache Kafka for inter-component communication, Apache Avro for data serialisation and Apache Spark for data analytics. A series of use cases are conducted for validation, which comprises workflow composition and semantic metadata management in academic and real-world fields of human activity recognition and land use monitoring from satellite images.