Technofist

Contact no: 9008001602

# A WORKFLOW MANAGEMENT SYSTEM FOR SCALABLE DATA MINING ON CLOUDS



Technofist,

YES Complex, 19/3&4, 2nd Floor, Dinnur Main Road, R.T.Nagar,Bangalore-560032

Ph:080-40969981, Website:www.technofist.com. E-mail:technofist.projects@gmail.com

**Abstract:**

The extraction of useful information from data is often a complex process that can be conveniently modeled as a data analysis workflow. When very large data sets must be analyzed and/or complex data mining algorithms must be executed, data analysis workflows may take very long times to complete their execution. Therefore, efficient systems are required for the scalable execution of data analysis workflows, by exploiting the computing services of the Cloud platforms where data is increasingly being stored. The objective of the paper is to demonstrate how Cloud software technologies can be integrated to implement an effective environment for designing and executing scalable data analysis workflows. We describe the design and implementation of the Data Mining Cloud Framework (DMCF), a data analysis system that integrates a visual workflow language and a parallel runtime with the Software-as-a-Service (SaaS) model. DMCF was designed taking into account the needs of real data mining applications, with the goal of simplifying the development of data mining applications compared to generic workflow management systems that are not specifically designed for this domain.



**Introduction:**

The past two decades have been characterized by an exponential growth of digital data production in many fields of human activities, from science to enterprise. Very large datasets are produced daily from sensors, instruments, mobile devices and computers, and are often stored in distributed repositories.

For example, astronomers analyze large image data that every day comes from telescopes and artificial satellites ; physicists must study the huge amount of data generated by particle accelerators to understand the laws of Universe; medical doctors and biologists collect huge amount of information about patients to search and try to understand the causes of diseases; sociologists analyze large social networks to find how users are influenced by others for various reasons. Such few examples demonstrate how the exploration and automated analysis of large datasets powered by computing capabilities are fundamental to advance our knowledge in many fields.