On the Use of Cloud Computing for Big Data Mining

Prof. Domenico Talia
DIMES, University of Calabria, Italy
talia@dimes.unical.it

Extended Abstract

Big data repositories are more and more massive and distributed, to manage them and make their contents useful, we need smart data analysis techniques and scalable architectures for extracting valuable information in reduced time. Cloud computing infrastructures offer an effective support for addressing both the computational and data storage needs of big data mining and parallel knowledge discovery applications. In fact, complex data mining tasks involve data- and compute-intensive algorithms that require large and efficient storage facilities together with high performance processors to get results in acceptable times.

We are addressing main topics and research issues on efficiently using Cloud computing platforms for implementing big data mining applications on large data sets. We present data mining techniques and frameworks designed for developing distributed data analytics applications on Clouds. These systems implement data set storage, analysis tools, data mining algorithms and knowledge models as single services that are combined through a visual programming interface in distributed workflows.

In particular, here we outline how to implement big data mining services on the Data Mining Cloud Framework, designed for developing and executing distributed data analytics applications as workflows of services. In this environment we use Big data, mining tools, data mining and machine learning algorithms and knowledge models that are implemented as services and can be composed in workflows through two programming interfaces: VS4Cloud and JS4Cloud designed for developing Big Data distributed workflows on Clouds. The main features of the programming interfaces are described and performance evaluation of some Big Data analysis applications is illustrated. The full implementation of the Data Mining Cloud Framework on Azure is presented and the main features of the graphical programming interface are described. Application design and execution of data analysis use cases are presented. Programming issues and research trends will be also outlined.

Clouds implement elastic services, scalable performance and scalable data storage used by a large and everyday increasing number of users and applications [1]. In fact, clouds enlarged the arena of distributed computing systems by providing advanced Internet services that complement and complete functionalities of distributed computing provided by the Web, Grid systems and peer-to-peer networks. In particular, most cloud computing applications use big data repositories stored within the cloud itself, so in those cases large datasets are analyzed with low latency to efficiently extract data analysis models.

Big data is a new and over-used term that refers to massive, heterogeneous, and often unstructured digital content that is difficult to process using traditional data management tools and techniques. The term includes the complexity and variety of data and data types, real-time data collection and processing needs, and the value that can be obtained by smart analytics. However we should recognize that data are not necessarily important per se but they become very important if we are able to extract value from them. That is if we can exploit them to make discoveries. The extraction of useful knowledge from big digital datasets requires smart and scalable analytics algorithms, services, programming tools, and applications. All these solutions need to find insights in big data will contribute to make them really useful for people.

The growing use of service-oriented computing is accelerating the use of cloud-based systems for scalable big data analysis. Developers and researchers are adopting the three main cloud models, software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS), to implement big data analytics solutions in the cloud [2]. Here we report on the main principles of the three models:

- Data analysis software as a service. Provides to end users data mining algorithms, data analysis suites or ready-to-use knowledge discovery tools offering them as Internet services that can be accessed directly through a Web browser.
- Data analysis platform as a service. Provides supporting platforms that developers can use to build their own data analytics applications or extend existing ones without concern about the underlying infrastructure or distributed architecture issues.
- Data analysis infrastructure as a service. Provides a set of virtualized resources that developers can assemble and use as a hardware/software infrastructure to store large data sets, run data mining applications and/or implement data analytics systems from scratch.

According to these approaches, data mining tasks and knowledge discovery applications are offered as high-level services available every time form everywhere. This methodology created a new way to delivery data analysis software that is called data analytics as a service (DAaaS).
Using the DAasS methodology we designed a cloud-based system, the Data Mining Cloud Framework (DMCF) [3], which supports three main classes of data analysis and knowledge discovery applications:

- Single-task applications, in which a single data mining task such as classification, clustering, or association rules discovery is performed on a given dataset;
- Parameter-sweeping applications, in which a dataset is analyzed by multiple instances of the same data mining algorithm with different parameters; and
- Workflow-based applications, in which knowledge discovery applications are specified as graphs linking together data sources, data mining tools, and data mining models.

DMCF includes a large variety of processing patterns to express knowledge discovery workflows as graphs whose nodes denote resources (datasets, data analysis tools, mining models) and whose edges denote dependencies among resources. A Web-based user interface allows users to compose their applications and submit them for execution to the Cloud platform, following the data analysis software as a service approach. Visual workflows can be programmed in DMCF through a language called VL4Cloud (Visual Language for Cloud), whereas script-based workflows can be programmed by JS4Cloud (JavaScript for Cloud), a JavaScript-based language for data analysis programming [4].

REFERENCES