Contribution to the European Data Forum 2015

Contributor and short CV

Domenico Talia is a full professor of computer engineering at the University of Calabria and a partner of two startups, Exeura and DtoK Lab. His research interests include parallel and distributed data mining, cloud computing, internet services, knowledge discovery, mobile computing, green computing systems, peer-to-peer systems, and parallel processing. Talia published ten books and more than 300 papers in archival journals such as CACM, Computer, IEEE TKDE, IEEE TSE, IEEE TSMC-B, IEEE Micro, ACM Computing Surveys, FGCS, Parallel Computing, IEEE Internet Computing and international conference proceedings. He is a member of the editorial boards of IEEE Transactions on Cloud Computing, the Future Generation Computer Systems journal, the International Journal on Web and Grid Services, the Scalable Computing: Practice and Experience journal, MultiAgent and Grid Systems: An International Journal, International Journal of Web and Grid Services, and the Web Intelligence and Agent Systems International journal. Talia has been a project for several international institutions such as the European Commission, Aeres in France, Austrian Science Fund, Croucher Foundation, and the Russian Federation Government. He served as a chair, organizer, or program committee member of several international conferences and gave many invited talks and seminars in conferences and schools. Talia is a member of the ACM and the IEEE Computer Society. Email: talia@dimes.unical.it

Type of the presentation proposed

Research contribution

Title of the presentation

Exploiting Cloud Services for Big Data Mining

Summary

In this presentation we discuss how to make data mining and knowledge discovery services scalable by exploiting the storage and computing facilities of Clouds. This will be done also presenting a data mining cloud framework designed for developing and executing distributed data analytics applications. In this framework we use data sets, analysis tools, data mining algorithms and knowledge models that are implemented as single services that can be combined through a visual programming interface in distributed workflows to be executed on clouds.

Extended Abstract

The amount of digital data is going to increase beyond any previous estimation and data stores and sources are more and more pervasive and distributed. Professionals

and scientists need advanced data analysis tools and services coupled with scalable architectures to support the extraction of useful information from big data repositories. Cloud computing systems 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.

As data sources became very large and pervasive, programming data analysis applications and services is a must to extract value and find useful insights in them. New ways to efficiently compose different distributed models and paradigms are needed and relationships between hardware resources and programming levels must be addressed. Users, professionals and scientists working in the area of big data need advanced data analysis tools and services coupled with scalable architectures to support the extraction of useful information from such massive repositories. Cloud computing platforms offer a real and scalable support for addressing both the computational and data storage needs of big data mining and parallel knowledge discovery applications. Complex data mining tasks involve data-intensive and compute-bound algorithms that require large and efficient storage facilities together with high performance processing units to get results in adequate times.

Cloud computing systems implement a computing model in which virtualized resources dynamically scalable are provided to users and developers as a service over the Internet. In fact, clouds implement scalable computing and storage delivery platforms that can be adapted to the needs of different classes of people and organizations by exploiting the Service Oriented (SOA) approach. The advent of clouds offered large facilities to many users that were unable to own their high-performance computing systems to run applications and services. In particular, big data analysis applications requiring access and manipulate very large datasets with complex mining algorithms will significantly benefit from the use of cloud platforms.

Although a few cloud-based analytics platforms are available today, current research work foresees that they will become common within a few years. Some current solutions are based on open source systems, such as Apache Hadoop and SciDB, while others are proprietary solutions provided by companies. As more such platforms emerge, researchers and professionals will port increasingly powerful data mining programming tools and strategies to the cloud to exploit complex and flexible software models such as the distributed workflow paradigm. The growing utilization of the service-oriented computing model could accelerate this trend.

This contribution discusses some approaches to make data analysis services scalable and introduces the Data Mining Cloud Framework (DMCF) designed for developing and executing distributed data analytics applications as workflows of services. In the DMCF environment developers can use datasets, analysis tools, data mining

algorithms and knowledge models that are implemented as single services. Each single service can be combined both through a visual and a script-programming interface in distributed workflows to be executed on clouds. The main features of the programming interface are described and performance figures of scalable data analysis applications are illustrated.

The big data analysis methodologies that can be adopted to implement data analysis tasks using the DMCF framework could be used in many application domains where data analysis techniques are useful to keep pace of the huge amount of data and of their complexity. We also discuss how clouds can be used to implement knowledge a set of specific data applications in scientific and business domains and outline some research and development issues to be further investigated.