

Making Knowledge Discovery Services Scalable on Clouds for Big Data Mining

Domenico Talia

DIMES-University of Calabria & DtoK Lab Srl

Rende (CS), Italy

`talia@dimes.unical.it`

Abstract— The amount of digital data is increasing 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. In this paper we introduce the topic and the main research issues. We discuss how to make knowledge discovery services scalable and present the Data Mining Cloud Framework designed for developing and executing distributed data analytics applications as workflows of services. In this environment 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. The main features of the programming interface are described and performance evaluation of knowledge discovery applications are reported.

Keywords— Big Data, Data mining, Cloud computing, Data mining cloud framework.

I. INTRODUCTION

Sources and repositories of digital data everyday increase beyond any previous estimate. Data centres and Web servers supporting Internet applications are uninterruptedly more and more pervasive and distributed. As data sources became very large and pervasive, programming data analysis applications and services is a must to 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.

This paper discusses how 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 through a visual or 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 described here, 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.

The rest of the paper is organized as follows. Section II introduces cloud computing concepts. Section III discusses how clouds can be used to implement knowledge discovery applications. Section IV gives an outline of the DMCF framework. Finally, Section V provides some final remarks and proposes some research issues to be further investigated.

II. CLOUD COMPUTING

The key features of clouds are: On-demand self-service, ubiquitous network access, location independent resource pooling, rapid elasticity, and pay per use. Since the cloud computing paradigm has been conceived several definitions have been given. Some definitions focus on on-demand dynamic provisioning of processing and storage resources, others emphasize the service-oriented model and the exploitation of virtualization [1]. The National Institute of

Standards and Technology (NIST) has given a complete reference definition. NIST defined clouds as follows: “*Cloud computing is a pay-per-use model for enabling available, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*” Moreover, according to NIST: “*Cloud model promotes availability and is comprised of five key characteristics, three delivery models, and four deployment models.*”

The delivery models of clouds are significant because they define three different types of cloud computing systems:

- *Infrastructure as a Service (IaaS)*. The capability provided to users is renting computing, storage, networks, and other computing resources where the user is able to deploy and run software, which can include operating systems and/or applications. A user does not manage or control the hardware cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly select networking components. Examples for commercial cloud infrastructures are Amazon EC2 and Rackspace.
- *Platform as a Service (PaaS)*. The functionality provided to users is deploying consumer-created applications onto the cloud infrastructure using programming languages, compilers and toolkits supported by the provider (e.g., Java, .Net, Oracle DB). A consumer does not manage or control the underlying cloud infrastructure, network, servers, operating environments, or storage, but she/he can control the deployed applications and possibly the configurations of the application hosting environment.
- *Software as a Service (SaaS)*. The capability provided to a consumer is to use the provider’s applications that run on a cloud infrastructure and are accessible from various client devices through a thin client interface such as a Web browser (e.g., Gmail, Dropbox). A consumer does not manage or control the underlying cloud infrastructure, network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings

Cloud computing is the a recent result of the advancement of several computer science technologies both from the hardware side, such as virtualization and multi-core architectures, and from the software side like cluster computing, Grid computing, Web services, service-oriented architectures, autonomic computing, and large-scale data storage. In particular, virtualization in cloud computing is the key element that separates system functionality and implementation from physical resources.

By exploiting virtualization techniques, a cloud infrastructure can be partitioned into several parallel virtual machines, dynamically configured according to user requirements and devoted to concurrently run independent applications. Virtualization splits applications from hardware and users from other users giving them the feeling that a

large-scale computing infrastructure is devoted to their applications by meeting a given quality of service (QoS). Virtualization is also used to isolate applications, so avoiding that failures in one of them do not result that other can fail too. Finally, virtualization is a technique to improve security and privacy of concurrent applications running on the same cloud.

As we can deduce from the previous descriptions, the cloud computing paradigm represents an advancement of the existing computing services available over the Internet. In particular, cloud infrastructures adopted the Web services paradigm for delivering new capabilities beyond the traditional Web capability.

Several companies set up large cloud facilities and built programming environments where developers can program applications as cloud software services. Just to mention a few examples, Google provides the AppEngine, Amazon on his EC2 and S3 cloud platforms implemented Elastic BeanStalk, Microsoft implemented .Net technology on Azure, and VMware implemented Cloud Foundry.

On the other side, the research community implemented open source software to be deployed and configured on private servers, computer farms or data centres for implementing private, public, community or hybrid cloud infrastructures or for inter-cloud computing facilities. Examples of these systems are Puppet, OpenQRM, OpenNebula, Eucalptus, and OpenStack. These open source software projects are also working to develop systems and services that allow cloud-to-cloud interoperability and federation.

III. BIG DATA ANALYSIS ON CLOUDS

The term big data designates massive, complex, and heterogeneous, digital data that is hard to process using traditional data management tools and techniques. Advanced data mining techniques and associated tools can efficiently support the extraction of information from huge and complex datasets that is useful in making informed decisions in several business and scientific application domains including tax payment collection, market analysis, economics, bio-sciences, and physics.

Although 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.

Analytics services can be implemented within each of the three fundamental Cloud service models [2]:

- *Data analysis as SaaS*, where a single well-defined data mining algorithm or a ready-to-use knowledge discovery tool is provided as an Internet service to end users, who may directly make use of it through a web browser.

- *Data analysis as PaaS*, where a supporting platform is provided to developers that have to build their own applications or extend existing ones. Developers can just focus on the definition of their data analysis applications without worrying about the underlying infrastructure or distributed computing issues.
- *Data analysis as IaaS*, where a set of virtualized resources (disks, cores, etc.) are provided to developers as a computing infrastructure to run their data mining applications or to implement their data analysis systems from scratch.

In all the scenarios listed above, cloud platforms play the role of hardware/software infrastructure provider, even if the SaaS and Paas modes make the infrastructure totally or partially transparent to end users.

IV. A FRAMEWORK FOR DATA ANALYSIS

To support the implementation of two of the data mining service models outlined above, has been implemented the Data Mining Cloud Framework (DMCF) [3]. It allows users to implement:

- *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 analysed by multiple instances of the same data mining algorithm with different parameters; and
- *Workflow-based applications*, in which knowledge discovery applications are developed as task graphs linking together data sources, data mining tools, and data mining models.

The DMCF framework includes also a programming interface and its services to support the composition and execution of workflow-based knowledge discovery applications. Workflows provide a paradigm that may encompass all the steps of discovery based on the execution of complex algorithms and the access and analysis of scientific data. In data-driven discovery processes, knowledge discovery workflows can produce results that can confirm real experiments or provide insights that cannot be achieved in laboratories.

Visual workflows in the framework are directed acyclic graphs whose nodes represent resources and whose edges represent the dependencies among the resources. Workflows include two types of nodes:

- *Data node*, which represents an input or output data element. Two subtypes exist: Dataset, which represents a data collection, and Model, which represents a model generated by a data analysis tool (e.g., a clustering dendrogram).
- *Tool node*, which represents a tool performing any kind of operation that can be applied to a data node (preprocessing, partitioning, classification, etc.).

The nodes can be connected with each other through direct edges, establishing specific dependency relationships among

them. When an edge is being created between two nodes, a label is automatically attached to it representing the kind of relationship between the two nodes. Data and Tool nodes can be added to the workflow singularly or in array form. A data array is an ordered collection of input/output data elements, while a tool array represents multiple instances of the same tool.

Figure 1. shows a sample data mining workflow composed of several sequential and parallel steps. It is just an example for presenting the main features of the visual programming interface of the DMCF [3].

The example workflow analyses a dataset by using n instances of a classification algorithm, which work on n partitions of the training set and generate the same number of knowledge models. By using the n generated models and the test set, n classifiers produce in parallel n classified datasets (n classifications). In the final step of the workflow, a voter generates the final classification by assigning a class to each data item, by choosing the class predicted by the majority of the models.

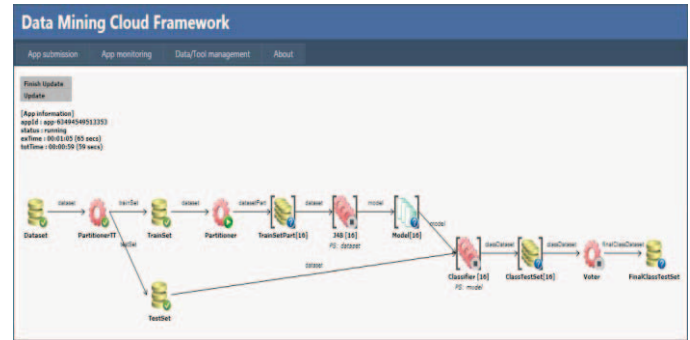


Figure 1. A parallel classification workflow during its execution.

Recently, we extended the DMCF to support also the design and execution of script-based data analysis workflows on Clouds [4]. To this end, we introduced a workflow language, named JS4Cloud, which extends JavaScript to support the definition of Cloud-based data analysis processes and the handling of data on the Cloud. JS4Cloud implements data-driven task parallelism. This means that, as soon as a task does not depend on any other task in the same workflow, the runtime asynchronously spawns it to the first available virtual machine. Figure 2. shows a JS4Cloud script implementing a parallel classification workflow equivalent to the visual workflow shown in Figure 1.

Several knowledge discovery applications have been recently implemented by using the DMCF software framework in different domains such as Internet monitoring, bioinformatics and smart cities [5].

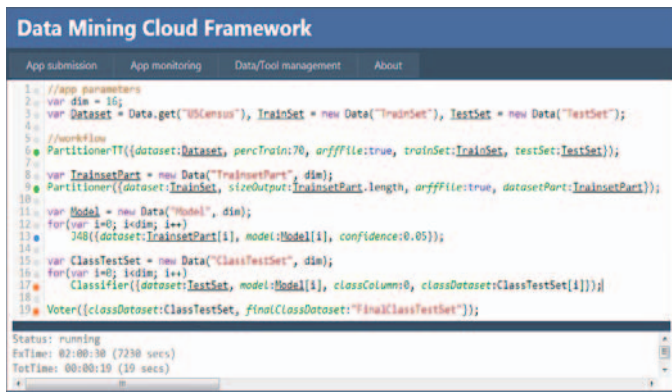


Figure 2. JS4Cloud workflow running in the DMCF's user interface.

V. CONCLUDING REMARKS AND RESEARCH TOPICS

The benefits of adopting cloud computing are clearly demonstrated by many reports from academic institutions and public administrations [6]. In particular, a 2011 survey by IDC for the European Commission shows that, as a result of the adoption of cloud computing, 80% of organisations reduce costs by 10-20%. Other benefits include enhanced mobile working (46%), productivity (41%), standardisation (35%), as well as new business opportunities (33%) and markets (32%) [7].

Cloud computing can provide scalable resources for big data mining and high-performance knowledge discovery applications. In fact, clouds offer large and efficient storage facilities with high performance processors to get results in reduced times. We introduced the key issue on developing knowledge discovery applications on clouds and sketched the min features of the Data Mining Cloud Framework designed for developing and running distributed data analytics applications as collections of services. In this framework, data sets, data mining algorithms and knowledge models are implemented as services that can be combined through a visual and a script-based programming interface to produce distributed workflows executed on Clouds.

Cloud-based data analytics requires high-level and technically easy-to-use design tools for programming large applications dealing with huge and/or distributed data sources. Despite the work done till today, further major efforts are needed in this area. Here we list some recommendations for fueling data analytics on Clouds and highlight a few key topics for further research and development:

- *High-level software tools and programming languages for big data analytics.* Large data analytics demands for further investigation towards higher and complex abstract structures to be included in big data programming tools. The MapReduce model is often used on clusters and clouds, but its expressiveness is limited and more research studies are needed for novel higher-level and scalable models and tools.
- *Data formats and tools interoperability and openness.*

Data and tool interoperability is a main issue in large-scale applications where many resources, data and computing nodes are used. Standard formats and models are needed to support interoperability and ease cooperation among teams using different data formats and tools.

- *Service-oriented workflows on multi-clouds.* The service-oriented paradigm allows large-scale distributed workflows to be running on heterogeneous platforms and the integration of software components developed by using different programming languages or tools. Web and cloud services are a paradigm that can help to handle integration in worldwide data analysis executed on multi-clouds, so this issue needs a deeper investigation.
- *Metadata tools, provenance and annotation mechanisms.* Metadata management tools are very useful to manage data according to their semantics. Data provenance is captured as a set of dependencies between data elements. It may be used for interpreting data and providing reproducible analysis. Research work towards novel techniques to annotate and manage provenance of data, to visualize and mine data before and during analysis is still needed.

These solutions, together with others addressing data privacy and security concerns, will foster and promote cloud-based data analytics and will help users - small research teams, start-ups, small enterprises - who aren't deeply skilled on cloud programming and management, benefiting from big data value by (semi-)automatic analysis. In his recent list about 13 new trends in big data and data science published in the Data Science Central online portal, Vincent Granville indicates that: "*High performance computing (HPC) could revolutionize the way algorithms are designed.*" In this scenario, the exploitation of cloud computing systems can act as accelerator of this trend for obtaining new scalable data analysis algorithms and applications.

REFERENCES

- [1] M. Armbrust, et al., "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50-58, April 2010.
- [2] D. Talia, Clouds for Scalable Big Data Analytics, *Computer*, 46(5), 98-101, May 2013
- [3] F. Marozzo, D. Talia, P. Trunfio, "A Cloud Framework for Big Data Analytics Workflows on Azure", In: *Cloud Computing and Big Data*, C. Catlett, W. Gentzsch, L. Grandinetti, G. Joubert, J. Vazquez-Poletti (Editors), IOS Press, Advances in Parallel Computing, vol. 23, pp. 182-191, 2013.
- [4] F. Marozzo, D. Talia, P. Trunfio, "Scalable script-based data analysis workflows on clouds", *Proc. of the 8th Workshop on Workflows in Support of Large-Scale Science (WORKS '13)*. ACM, New York, NY, USA, pp. 124-133, 2013.
- [5] A. Altomare, E. Cesario, C. Comito, F. Marozzo, D. Talia, "Using Clouds for Smart City Applications", *Proc. of the 5th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2013)*, Bristol, UK, 2013.
- [6] Cloud Computing Expert Group, "The Future of Cloud Computing," Report from European Commission, January 2010.
- [7] IDC, "Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Take-up", 2012.