2581-4575

International Journal For Recent Developments in Science & Technology

Crossref

A Peer Reviewed Research Journal

# KNOWLEDGE DISCOVERY SERVICES SCALABLE ON CLOUD DATA

DR. T. PREM CHANDER

Associate Professor, ISL Engineering College, Hyderabad, India

Email: tudiprem@gmail.com

**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. 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 discuss how to make security services scalable and present the Data Mining Cloud Framework designed for developing and executing distributed data analytics applications as workflows of services.

***Keywords*— 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.

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 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.

## II. CLOUD COMPUTING

The key features of clouds are:

1. On-demand self-service
2. Ubiquitous network access
3. Location independent resource pooling
4. Rapid elasticity and
5. 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 and data hiding is concerned with the data sequence. The extraction of data is information oriented with a least consideration of level of computed algorithm. The complex data computing algorithm consisted with a resultant bound upper and to a sequence of massive repositories with the data still exits.

   The following are the sequential data centered task at every level of access and can be taken to over data exchange:

   1. Centric data download
   2. Transactional data overload.

Standards and Technology has given a complete reference definition. DATA CENTRES 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 DATA CENTRES:

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,
- *Platform as a Service* (*PaaS*). The functionality provided to users is deploying consumer-created applications onto the cloud infrastructure using programming languages.
- *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. 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 has data centered data and can store only one transational computed data field. The presence of data is most important in multi streams.

The following are the occurances sequences:

1. Computed node sequence.
2. Data unliked nature of occurance.
3. Getting data at a stream of data flow in each node.

large-scale computing infrastructure is devoted to their applications by meeting a given quality of service. Virtualization is also used to isolate applications, so avoiding that failures in one of them do not result that other can fail too.

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.

## III.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 : *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.

Data Mining Cloud Framework

1. It allows users to implement:
2. *Single-task applications*
3. *Parameter-sweeping application*
4. *Workflow-based applications*

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:

5. *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).

6. *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 of several sequential and parallel steps. It is just an example for presenting the main features of the visual programming interface of the DMCF .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.Recently, we extended the DMCF to support also the design and execution of script-based data analysis workflows on Clouds. Several knowledge discovery applications have been recently implemented by 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:

7. *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- using the DMCF software framework in different domains such as Internet monitoring, bioinformatics and smart cities and its collection of data at one instance.

## IV. CONCLUSION.

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 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.

## 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.
F. Marozzo, D. Talia, P. Trunfio, “A Cloud
Framework for Big Data
Analytics Workflows on Azure”, In: *Cloud*
[3] *Computing and Big Data*,
C. Catlett, W. Gentzsch, L. Grandinetti, G.
Joubert, J. Vazquez-Poletti (Editors), IOS
[4] Press, Advances in Parallel Computing, vol. 23,
pp. 182- 191, 2013.
F. Marozzo, D. Talia, P. Trunfio, “Scalable
[5] 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.
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.

level and scalable models and tools.

- *Data formats and tools interoperability and openness*.

Cl Computing Expert Group, "The Future of Cloud
o Computing,"
u Report from European Commission, January
d 2010.

[6] IDC, "Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Take-up", 2012.