Cloud computing for Big Data analysis

Fabrizio Marozzo, Loris Belcastro

Definitions

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell and Grance (2011)).

Overview

In the last decade the ability to produce and gather data has increased exponentially. For example, huge amounts of digital data are generated by and collected from several sources, such as sensors, web applications and services. Moreover, thanks to the growth of social networks (e.g., Facebook, Twitter, Pinterest, Instagram, Foursquare, etc.) and the widespread diffusion of mobile phones every day millions of people share information about their interests and activities. The amount of data generated, the speed at which it is produced, and its heterogeneity in terms of format, represent a challenge to the current storage, process and analysis capabilities. Those data volumes, commonly referred as Big Data, can be exploited to extract useful information and to produce helpful knowledge for science, industry, public services and in general for humankind.

To extract value from such data, novel technologies and architectures have been developed by data scientists for capturing and analyzing complex and/or high velocity data. In general, the process of knowledge discovery from Big Data is not so easy, mainly due to data characteristics, as size, complexity and variety, that require to address several issues. To overcame these problems and get valuable information and knowledge in shorter time, high performance and scalable computing systems are used in combination with data and knowledge discovery techniques.

In this context, Cloud computing has emerged as an effective platform to face the challenge of extracting knowledge from Big Data repositories in limited time, as well as to provide an effective and efficient data analysis environment for researchers and companies. From a client perspective, the Cloud is an abstraction for remote, infinitely scalable provisioning of computation and storage resources (Talia et al (2015)). From an implementation point of view, Cloud systems are based on large sets of computing resources, located somewhere "in the Cloud", which are allocated to applications on demand (Barga et al (2011)).

Fabrizio Marozzo e-mail: fmarozzo@dimes.unical.it and Loris Belcastro e-mail: lbelcastro@dimes.unical.it
DIMES, University of Calabria, Rende (Italy)

Thus, Cloud computing can be defined as a distributed computing paradigm in which all the resources, dynamically scalable and often virtualized, are provided as services over the Internet. As defined by NIST (National Institute of Standards and Technology) (Mell and Grance (2011)), Cloud computing can be described as: "*A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction*". From the NIST definition, we can identify five essential characteristics of Cloud computing systems, which are: *i*) on-demand self-service, *ii*) broad network access, *iii*) resource pooling, *iv*) rapid elasticity, and *v*) measured service.

Cloud computing vendors provide their services according to three main distribution models:

- *Software as a Service (SaaS)*, in which software and data are provided through Internet to customers as ready-to-use services. Specifically, software and associated data are hosted by providers, and customers access them without need to use any additional hardware or software.
- *Platform as a Service (PaaS)*, in an environment including databases, application servers, development environment for building, testing and running custom applications. Developers can just focus on deploying of applications since Cloud providers are in charge of maintenance and optimization of the environment and underlying infrastructure.
- *Infrastructure as a Service (IaaS)*, that is an outsourcing model under which customers rent resources like CPUs, disks, or more complex resources like virtualized servers or operating systems to support their operations. Compared to the PaaS approach, the IaaS model has a higher system administration costs for the user; on the other hand, IaaS allows a full customization of the execution environment.

Key Research Findings

Most available data analysis solutions today are based on open source frameworks, such as Hadoop[1] and Spark[2], but there are also some proprietary solutions proposed by big companies (e.g., IBM, Kognitio).

Concerning the distribution models of Cloud services, the most common ones for providing Big Data analysis solutions are PaaS and SaaS. Usually, IaaS is not used for high-level data analysis applications but mainly to handle the storage and computing needs of data analysis processes. In fact, IaaS is the more expensive distribution model, because it requires a greater investment of IT resources. On the contrary, PaaS is widely used for Big Data analysis, because it provides data analysts with tools, programming suites, environments, and libraries ready to be built, deployed and run on the Cloud platform. With the PaaS model users do not need to care about configuring and scaling the infrastructure (e.g., a distributed and scalable Hadoop system), because the Cloud vendor will do that

[1] https://hadoop.apache.org/

[2] https://spark.apache.org/

for them. Finally, the SaaS model is used to offer complete Big Data analysis applications to end users, so that they can execute analysis on large and/or complex datasets by exploiting Cloud scalability in storing and processing data.

As outlined in Li et al (2010), users can access Cloud computing services using different client devices, Web browsers and desktop/mobile applications. The business software and user data are executed and stored on servers hosted in Cloud data centers, which provide storage and computing resources. Such resources include thousands of servers and storage devices connected each other through an intra-Cloud network.

Several technologies and standards are used by the different components of the architecture. For example, users can interact with Cloud services through SOAP-based or RESTful Web services (Richardson and Ruby (2008)) and Ajax technologies, which let Cloud services to have look and interactivity equivalent to those provided by desktop applications.

Developing Cloud-based Big Data analysis applications may be a complex task, with specific issues that go beyond those of stand-alone application programming. For instance, Cloud programming must deal with deployment, scalability and monitoring aspects that are not easy to handle without the use of ad-hoc environments (Talia et al (2015)). In fact, to simplify the development of Cloud applications, specific development environments are often used. Some of the most representative Cloud computing development environments currently in use can be classified into four types:

- *Integrated development environments*, which are used to code, debug, deploy and monitor Cloud applications that are executed on a Cloud infrastructure, such as Eclipse, Visual Studio and IntelliJ.
- *Parallel-processing development environments*, which are used to define parallel applications for processing large amount of data that are run on a cluster of virtual machines provided by a Cloud infrastructure (e.g., Hadoop and Spark).
- *Workflow development environments*, which are used to define workflow-based applications that are executed on a Cloud infrastructure, such as Swift and DMCF.
- *Data-analytics development environments*, which are used to define data analysis applications through machine learning and data mining tools provided by a Cloud infrastructure. Some examples are Azure ML and BigML.

The programming model is a key factor to be considered for exploiting the powerful features of Cloud computing. MapReduce (Dean and Ghemawat (2004)) is widely recognized as one of the most important programming models for Cloud computing environments, being it supported by Google and other leading Cloud providers such as Amazon, with its Elastic MapReduce service, and Microsoft, with its HDInsight, or on top of private Cloud infrastructures such as OpenNebula, with its Sahara service. Hadoop is the best-known MapReduce implementation and it is commonly used to develop parallel applications that analyze big amounts of data on Clouds. In fact, Hadoop-ecosystem is undoubtedly one of the most complete solution for data analysis problem, but

at the same time it is thought for high skilled users.

On the other hand, many other solutions are designed for low-skilled users or for low-medium organizations that do not want to spend resources in developing and maintaining enterprise data analysis solutions. Two representative examples of such data analysis solutions are Microsoft Azure Machine Learning and Data Mining Cloud Framework.

Microsoft Azure Machine Learning (Azure ML)[3] is a SaaS for the creation of machine learning workflows. It provides a very high-level of abstraction, because a programmer can easily design and execute data analytics applications by using simple drag-and-drop web interface and exploiting many built-in tools for data manipulation and machine learning algorithms.

The Data Mining Cloud Framework (DMCF) (Marozzo et al (2015)) is a software system developed at University of Calabria for allowing users to design and execute data analysis workflows on Clouds. DMCF supports a large variety of data analysis processes, including single-task applications, parameter sweeping applications, and workflow-based applications. A workflow in DMCF can be developed using a visual- or a script-based language. The visual language, called VL4Cloud (Marozzo et al (2016)), is based on a design approach for end users having a limited knowledge of programming paradigms. The script-based language, called JS4Cloud (Marozzo et al (2015)), provides a flexible programming paradigm for skilled users who prefer to code their workflows through scripts.

[3] https://azure.microsoft.com/services/machine-learning-studio/

Other solutions have been created mainly for scientific research purposes and, for this reason, they are poorly used for developing business applications (e.g., E-Science Central, COMPSs, and Sector/Sphere).

E-Science Central (e-SC) (Hiden et al (2013)) is a Cloud-based system that allows scientists to store, analyze and share data in the Cloud. It provides a user interface that allows programming visual workflows in any Web browser.

e-SC is commonly used to provide a data analysis back end to standalone desktop or Web applications. To this end, the e-SC API provides a set of workflow control methods and data structures. In the current implementation, all the workflow services within a single invocation of a workflow execute on the same Cloud node.

COMPSs (Lordan et al (2014)) is a programming model and an execution runtime, whose main objective is to ease the development of workflows for distributed environments, including private and public Clouds. With COMPSs, users create a sequential application and specify which methods of the application code will be executed remotely. Providing an annotated interface where these methods are declared with some metadata about them and their parameters does this selection. The runtime intercepts any call to a selected method creating a representative task and finding the data dependencies with all the previous ones that must be considered along the application run.

Sector/Sphere (Gu and Grossman (2009)) is an open source Cloud framework designed to implement data analysis applications involving large, geographically distributed datasets. The framework includes its own storage

and compute services, called Sector and Sphere respectively, which allow to manage large dataset with high performance and reliability.

Examples of Application

Cloud computing has been used in many scientific fields, such as astronomy, meteorology, social computing, and bioinformatics, which are greatly based on scientific analysis on large volume of data. In many cases, developing and configuring Cloud-based applications requires an high level of expertise, which is a common bottleneck in the adoption of such applications by scientists.

Many solutions for Big Data analysis on Clouds have been proposed in bioinformatics, such as: Myrna (Langmead et al (2010)) is a Cloud system that exploits MapReduce for calculating differential gene expression in large RNA-seq datasets;

Wang et al (2015) propose a heterogeneous Cloud framework exploiting MapReduce and multiple hardware execution engines on FPGA to accelerate the genome sequencing applications.

Cloud computing has been also used for executing complex Big Data mining applications. Some examples are: Agapito et al (2013) perform an association rule analysis between genome variations and clinical conditions of a large group of patients; Altomare et al (2017) propose a Cloud-based methodology to analyze data of vehicles in a wide urban scenario for discovering patterns and rules from trajectory; Kang et al (2012) present a library for scalable graph mining in the Cloud that allows to find patterns and anomalies in massive, real-world graphs; Belcastro et al (2016) propose a model for predicting flight delay according to weather conditions.

Several other works exploited Cloud computing for conducting data analysis on large amount of data gathered from social networks. Some examples are:

You et al (2014) propose a social sensing data analysis framework in Clouds for smarter cities, especially to support smart mobility applications (e.g., finding crowed areas where more transportation resources need to be allocated); Belcastro et al (2017) present a Java library, called ParSoDA (Parallel Social Data Analytics), which can be used for developing social data analysis applications.

Future Directions for Research

Some of most important research trends and issues to be addressed in Big Data analysis and Cloud systems for managing and mining large-scale data repositories are:

- *Data-intensive computing*. The design of data-intensive computing platforms is a very significant research challenge with the goal of building computers composed of a large number of multi-core processors. From a software point of view, these new computing platforms open big issues and challenges for software tools and runtime systems that must be able to manage a high degree of parallelism and data locality. In addition, to provide efficient methods for storing, accessing and communicating data, intelligent techniques for data analysis and scalable software

architectures enabling the scalable extraction of useful information and knowledge from data, are needed.

- *Massive social network analysis*. The effective analysis of social network data on a large scale requires new software tools for real-time data extraction and mining, using Cloud services and high-performance computing approaches (Martin et al (2016)). Social data streaming analysis tools represent very useful technologies to understand collective behaviors from social media data. New approaches to data exploration and model visualization are necessary taking into account the size of data and the complexity of the knowledge extracted.
- *Data quality and usability*. Big Data sets are often arranged by gathering data from several heterogeneous and often not well-known sources. This leads to a poor data quality that is a big problem for data analysts. In fact, due to the lack of a common format, inconsistent and useless data can be produced as a result of joining data from heterogeneous sources. Defining some common and widely adopted format would lead to data that are consistent with data from other sources, that means high quality data.
- *In-memory analysis*. Most of the data analysis tools access data sources on disks while, differently from those, in-memory analytics access data in main memory (RAM). This approach brings many benefits in terms of query speed up and faster decisions. In-memory databases are, for example, very effective in real-time data analysis, but they require high-performance hardware support and fine-grain parallel algorithms (Tan et al (2015)). New 64-bit operating systems allow to address memory up to one terabyte, so making realistic to cache very large amount of data in RAM. This is why this research area has a strategic importance.
- *Scalable software architectures for fine grain in-memory data access and analysis*. Exascale processors and storage devices must be exploited with fine-grain runtime models. Software solutions for handling many cores and scalable processor-to-processor communications have to be designed to exploit exascale hardware (Mavroidis et al (2016)).

References

Agapito G, Cannataro M, Guzzi PH, Marozzo F, Talia D, Trunfio P (2013) Cloud4snp: Distributed analysis of snp microarray data on the cloud. In: Proc. of the ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics 2013 (ACM BCB 2013), ACM Press, Washington, DC, USA, p 468, iSBN 978-1-4503-2434-2

Altomare A, Cesario E, Comito C, Marozzo F, Talia D (2017) Trajectory pattern mining for urban computing in the cloud. Transactions on Parallel and Distributed Systems 28(2):586–599, iSSN:1045-9219

Barga R, Gannon D, Reed D (2011) The client and the cloud: Democratizing research computing. Internet Computing, IEEE 15(1):72–75, DOI 10.1109/MIC.2011.20

Belcastro L, Marozzo F, Talia D, Trunfio P (2016) Using scalable data mining for predicting flight delays. ACM Transactions on Intelligent Systems and Technology (ACM TIST) To appear

Belcastro L, Marozzo F, Talia D, Trunfio P (2017) A parallel library for social media analytics. In: The 2017 International Conference on High Performance Computing & Simulation (HPCS 2017), Genoa, Italy, pp 683–690, iSBN 978-1-5386-3250-5

Cesario E, Iannazzo AR, Marozzo F, Morello F, Riotta G, Spada A, Talia D, Trunfio P (2016) Analyzing social media data to discover mobility patterns at expo 2015: Methodology and results. In: The 2016 International Conference on High Performance Computing & Simulation (HPCS 2016), Innsbruck, Austria, pp 230–237, iSBN: 978-1-5090-2088-1

Dean J, Ghemawat S (2004) Mapreduce: Simplified data processing on large clusters. In: Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6, Berkeley, USA, OSDI'04, pp 10–10

Gu Y, Grossman RL (2009) Sector and sphere: the design and implementation of a high-performance data cloud. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 367(1897):2429–2445

Hiden H, Woodman S, Watson P, Cala J (2013) Developing cloud applications using the e-science central platform. Phil Trans R Soc A 371(1983):20120,085

Jourdren L, Bernard M, Dillies MA, Le Crom S (2012) Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics 28(11):1542–1543

Kang U, Chau DH, Faloutsos C (2012) Pegasus: Mining billion-scale graphs in the cloud. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 5341–5344, DOI 10.1109/ICASSP.2012.6289127

Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson KE (2012) Cloud biolinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC bioinformatics 13(1):42

Langmead B, Hansen KD, Leek JT (2010) Cloud-scale rna-sequencing differential expression analysis with myrna. Genome biology 11(8):R83

Li A, Yang X, Kandula S, Zhang M (2010) Cloudcmp: comparing public cloud providers. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, ACM, pp 1–14

Lordan F, Tejedor E, Ejarque J, Rafanell R, lvarez J, Marozzo F, Lezzi D, Sirvent R, Talia D, Badia R (2014) Servicess: An interoperable programming framework for the cloud. Journal of Grid Computing 12(1):67–91

Marozzo F, Talia D, Trunfio P (2015) Js4cloud: Script-based workflow programming for scalable data analysis on cloud platforms. Concurrency and Computation: Practice and Experience 27(17):5214–5237

Marozzo F, Talia D, Trunfio P (2016) A workflow management system for scalable data mining on clouds. IEEE Transactions On Services Computing

Martin A, Brito A, Fetzer C (2016) Real-time social network graph analysis using streammine3g. In: Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems, ACM, New York, NY, USA, DEBS '16, pp 322–329

Mavroidis I, Papaefstathiou I, Lavagno L, Nikolopoulos DS, Koch D, Goodacre J, Sourdis I, Papaefstathiou V, Coppola M, Palomino M (2016) Ecoscale: Reconfigurable computing and runtime system for future exascale systems. In: 2016 Design, Automation Test in Europe Conference Exhibition (DATE), pp 696–701

Mell PM, Grance T (2011) Sp 800-145. the nist definition of cloud computing. Tech. rep., National Institute of Standards & Technology, Gaithersburg, MD, United States

Richardson L, Ruby S (2008) RESTful web services. "O'Reilly Media, Inc."

Talia D, Trunfio P, Marozzo F (2015) Data Analysis in the Cloud. Elsevier, iSBN 978-0-12-802881-0

Tan KL, Cai Q, Ooi BC, Wong WF, Yao C, Zhang H (2015) In-memory databases: Challenges and opportunities from software and hardware perspectives. SIGMOD Rec 44(2):35–40

Wang C, Li X, Chen P, Wang A, Zhou X, Yu H (2015) Heterogeneous cloud framework for big data genome sequencing. IEEE/ACM Trans Comput Biol Bioinformatics 12(1):166–178, DOI 10.1109/TCBB.2014.2351800

You L, Motta G, Sacco D, Ma T (2014) Social data analysis framework in cloud and mobility analyzer for smarter cities. In: Service Operations and Logistics, and Informatics (SOLI), 2014 IEEE International Conference on, pp 96–101