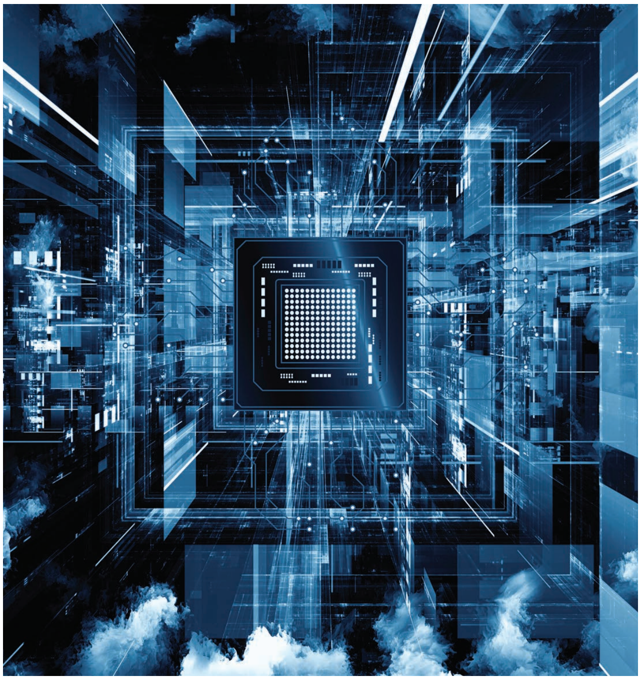


Clouds for Scalable Big Data Analytics

Domenico Talia
University of Calabria, Italy

Extracting useful knowledge from huge digital datasets requires smart and scalable analytics services, programming tools, and applications.

The proliferation of data warehouses, webpages, audio and video streams, tweets, and blogs is generating a massive amount of complex and pervasive digital data. Efficient means are now available for creating, storing, and sharing this information, which also fuels data growth. However, extracting useful knowledge from huge digital datasets requires smart and scalable analytics services, programming tools, and applications.

Big data analytics use compute-intensive data mining algorithms that require efficient high-performance processors to produce timely results. Cloud computing infrastructures can serve as an effective platform for addressing both the computational and data storage needs of big data analytics applications.

Much big data already resides in the cloud, and this trend will increase in the future. For example, IT research and advisory firm Gartner estimates that, by 2016, more than half of large companies' data will be stored in the cloud (http://tinyurl.com/boltvbo). This trend requires that clouds become the infrastructure for implementing pervasive and scalable data analytics platforms.

Coping with and gaining value from cloud-based big data requires novel software tools and innovative analytics techniques.

TOWARD CLOUD-BASED BIG DATA ANALYTICS

Big data refers to massive, heterogeneous, and often unstructured digital content that is difficult to process using traditional data management tools and techniques. The term encompasses 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.

Advanced data mining techniques and associated tools can help extract information from large, complex datasets that is useful in making informed decisions in many business and scientific applications including tax payment collection, market sales, social studies, biosciences, and high-energy physics. Combining big data analytics and knowledge discovery techniques with scalable computing systems will produce new insights in a shorter time.

Although few cloud-based analytics platforms are available today, current research work anticipates that they will become common within a few years. Some current solutions are based on open

Published by the IEEE Computer Society

source systems such as Apache Hadoop and SciDB, while others are proprietary solutions provided by companies such as Google, IBM, EMC, BigML, Splunk Storm, Kognitio, and InsightsOne.

As more such platforms emerge, researchers 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 use of service-oriented computing could accelerate this trend (http://tinyurl.com/d26o2j5).

DATA ANALYTICS SERVICE MODELS

Developers and researchers can adopt the software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS) models to implement big data analytics solutions in the cloud.

The SaaS model offers complete big data analytics applications to end users, who can exploit the cloud's scalability in both data storage and processing power to execute analysis on large or complex datasets. The PaaS model provides data analytics programming suites and environments in which data mining developers can design scalable analytics services and applications. Researchers can exploit the IaaS model to compose a set of virtualized hardware and software resources for running data analysis frameworks or applications.

COLUMN CONTRIBUTIONS

We welcome short articles (1,500 to 2,000 words) for publication in the Cloud Cover column that address the questions outlined in *Computer's* January 2013 issue (S. Murugesan, "Cloud Computing: The New Normal?," pp. 77-79). Submit your ideas for advancing the technology or share your experiences in harnessing the cloud at cloudcover@computer.org.

As Table 1 shows, developers can implement big data analytics services within each of these three models:

- *data analytics software as a service*—provides a well-defined data mining algorithm or ready-to-use knowledge discovery tool as an Internet service to end users, who can access it directly through a Web browser;
- *data analytics platform as a service*—provides a supporting platform that developers can use to build their own data analytics applications or extend existing ones without concern about the underlying infrastructure or distributed computing issues; and
- *data analytics infrastructure as a service*—provides a set of virtualized resources that developers can use as a computing infrastructure to run data mining applications or to implement data analytics systems from scratch.

End users whose goal is to perform complex data analysis can apply the recently implemented Data Mining Cloud Framework (http://tinyurl.com/c4b4f5k) as a high-level PaaS programming environment and create a set of SaaS suites for big data analytics. With this approach, users need not be concerned about cloud platform or application programming details.

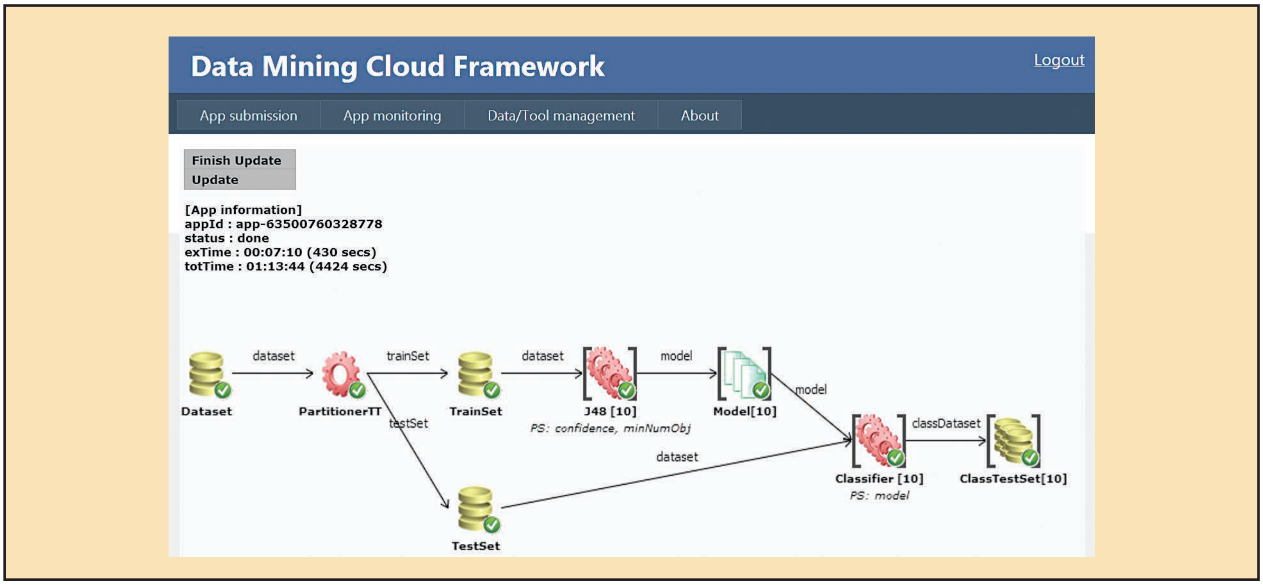
BIG DATA ANALYTICS WORKFLOWS

Developers can use workflows, which consist of complex graphs of many concurrent tasks, to address the complexity of scientific and business applications. This approach supports data analytics design by providing a paradigm that encompasses all the steps of data analytics, from data access and filtering to data mining and sharing produced knowledge.

Workflow-based data mining frameworks that run on cloud platforms and use a service-oriented approach offer a flexible programming model, distributed task interoperability, and execution scalability that reduces data analytics completion time. Application developers can design

Table 1. Cloud-based data analytics services.

Cloud service model	Features	Users
Data analytics software as a service	A single and complete data mining application or task (including data sources) offered as a service	End users, analytics managers, data analysts
Data analytics platform as a service	A data analysis suite or framework for programming or developing high-level applications, hiding the cloud infrastructure and data storage	Data mining application developers, data scientists
Data analytics infrastructure as a service	A set of virtualized resources provided to a programmer or data mining researcher for developing, configuring, and running data analysis frameworks or applications	Data mining programmers, data management developers, data mining researchers



[App information]
appId : app-63500760328778
status : done
exTime : 00:07:10 (430 secs)
totTime : 01:13:44 (4424 secs)

Figure 1. Data analysis workflow application designed using the Data Mining Cloud Framework's graphical programming interface.

data analysis tasks, scientific computation methods, and complex simulation techniques as workflows that integrate single Web services and execute them concurrently on virtual machines in the cloud.

Figure 1 shows a data analysis workflow application designed using the Data Mining Cloud Framework's graphical programming interface recently developed in our laboratory (http://tinyurl.com/crnork2). Data sources and tools such as data mining algorithms, filters, and data splitters are connected through direct edges that define specific dependency relationships among them.

When creating an edge between two nodes, the system automatically attaches a label to it that represents the relationship between them. To ease workflow composition and allow users to monitor its execution, each resource icon has an associated tag—the checkmarks in Figure 1—representing the status of a corresponding resource.

The experimental results of a set of studies using the framework to analyze genomics, network intrusion, and bioinformatics data demonstrated its effectiveness, as well as the linear scalability achieved through concurrent execution of the workflow tasks on a pool of virtual servers (http://tinyurl.com/c4b4f5k).

Current research focuses on the workflow composition interface, with the aim of extending supported design patterns such as conditional branches and iterations and evaluating its functionality and performance during the design and execution of complex data mining workflows on large datasets in the cloud.

RESEARCH RECOMMENDATIONS

Cloud-based data analytics requires high-level, easy-to-use design tools for programming large applications dealing with huge, distributed data sources. This necessitates further research and development in several key areas.

- *Programming abstracts for big data analytics*. Big data analytics programming tools require novel complex abstract structures. The MapReduce model is often used on clusters and clouds, but more research is needed to develop scalable higher-level models and tools.
- *Data and tool interoperability and openness*. Interoperability is a main issue in large-scale applications that use resources such as data and computing nodes. Standard formats and models are needed to support interoperability and ease cooperation among teams using

CLOUD COMPUTING SPECIAL TECHNICAL COMMUNITY

The CS Cloud Computing Special Technical Community (CS CCSTC) focuses on cloud activities across the IEEE Computer Society, involving both CS members and nonmembers. The STC's work is complementary to the IEEE Cloud Computing Initiative (IEEE CCI), a three-year project to promote cloud efforts across IEEE.

For details or to join the STC, visit www.computer.org/cc.

different data formats and tools.

- *Integration of big data analytics frameworks*. The service-oriented paradigm allows running large-scale distributed workflows on heterogeneous platforms along with software components developed using different programming languages or tools. The Web and cloud services paradigms can help manage worldwide integration of multiple data analytics frameworks.
- *Data provenance and annotation mechanisms*. Provenance is captured as a set of dependencies between elements that researchers can use to interpret data and provide reproducible analysis. Research is needed to develop innovative techniques for visualizing and mining provenance data.

These solutions, together with others addressing data privacy and security concerns, will promote cloud-based data analytics in large companies, and eventually will benefit users such as independent research teams, start-ups, and small enterprises that aren't deeply skilled in cloud programming and management.

Advancing the cloud from a computation and data management infrastructure to a pervasive and scalable data analytics platform requires new models, tools, and technologies that support the implementation of dynamic data analysis algorithms. Using a synergic approach that integrates the use of clouds and data analysis techniques to investigate key issues will help researchers achieve this goal. C

Domenico Talia *is a professor of computer engineering at the University of Calabria, Italy, and the director of ICAR-CNR (Institute for High-Performance Computing and Networking—Italian National Research Council). Contact him at talia@deis.unical.it.*

Editor: San Murugesan, BRITE Professional Services, Australia; san1@internode.net or Twitter @santweets.