The Crime and Corruption Observatory is part of a broader European project called FuturICT, which was the first of six "FET Flagship"pilots selected by the European Commission as part of the Framework 7 Programme.

The mission of FuturICT is to unleash the power of information for a sustainable future: a Living Earth Simulator will be built to understand and manage complex social systems, with a focus on sustainability and resilience. Within this framework, the Crime and Corruption Observatory will identify the underlying economical, social and cultural mechanisms that influence illegal phenomena, in order to control them at European level.

Addressing important issues, such as fighting against terrorism and organized crime, fraud detection, and maintaining internal and external security, certainly requires the use of modern technology, but this is not enough. Data must be transformed into information and then into knowledge, to reveal the real meaning of the billions of bits gathered worldwide. For this reason, behind Big Data lie Big Questions: the Crime and Corruption Observatory will identify fundamental issues about the dynamics of crime and their implications, developing solutions and innovative theories at the same time.

From BigData to virtual models of our society (image by courtesy of FuturICT)

Data technology will thus become both responsive and responsible: it will provide not only practical answers but also reliable theoretical tools, since Big Data should always have underlying Big Questions.

Link: http://www.futurict.eu

Please contact:
Giulia Bonelli
ISTC-CNR, Italy
Tel: +39 338 8689020
E-mail: giulia.bonelli@istc.cnr.it

Mario Paolucci, ISTC-CNR, Italy
E-mail: mario.paolucci@istc.cnr.it

Rosaria Conte, ISTC-CNR, Italy
E-mail: rosaria.conte@istc.cnr.it

Managing Big Data through Hybrid Data Infrastructures

by Leonardo Candela, Donatella Castelli and Pasquale Pagano

Long-established technological platforms are no longer able to address the data and processing requirements of the emerging data-intensive scientific paradigm. At the same time, modern distributed computational platforms are not yet capable of addressing the global, elastic, and networked needs of the scientific communities producing and exploiting huge quantities and varieties of data. A novel approach, the Hybrid Data Infrastructure, integrates several technologies, including Grid and Cloud, and promises to offer the necessary management and usage capabilities required to implement the 'Big Data' enabled scientific paradigm.

A recent study, promoted by The Royal Society of London in cooperation with Elsevier, reviewed the changing patterns of science highlighting that science is increasingly a global, multidisciplinary and networked effort performed by scientists that dynamically collaborate to achieve specific objectives. The same study also indicated that data-intensive science is gaining momentum in many domains. Large-scale datasets come in all forms and shapes from huge international experiments to cross-laboratory, single laboratory, or even from a multitude of individual observations.

The management and processing of such datasets is beyond the capacity of traditional technological approaches based on local, specialized data facilities. They require innovative solutions able to simultaneously address the needs imposed by multidisciplinary collaborations and by the new data-intensive pattern. These needs are characterized by the well known three V's: (i) Volume – data dimension in terms of bytes is huge, (ii) Velocity – data collection, processing and consumption is demanding in terms of speed, and (iii) Variety – data heterogeneity, in terms of data types and data sources requiring integration, is high.

Recent approaches, such as Grid and Cloud Computing, can only partially satisfy these needs. Grid Computing was initially conceived as a technological platform to overcome the limitations in volume and velocity of single laboratories by sharing and re-using computational and storage resources across laboratories. It offers a valid solution in specific scientific domains such as High Energy Physics. However, Grid Computing does not handle

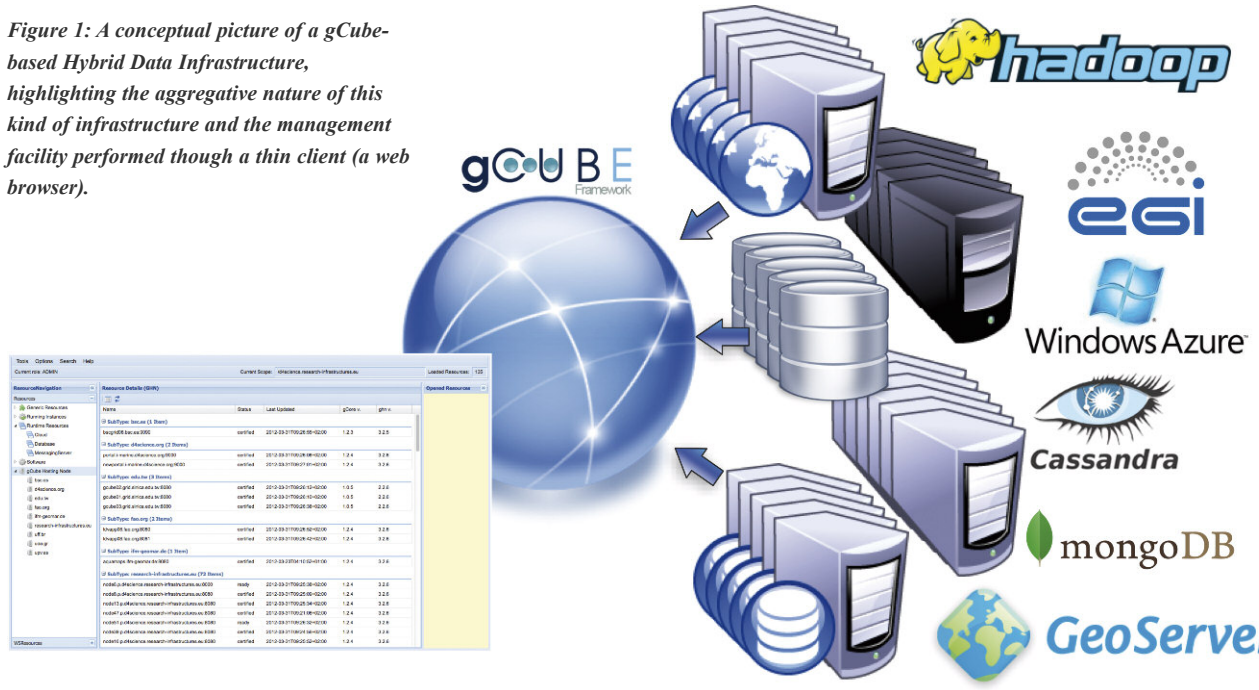

Figure 1: A conceptual picture of a gCube-based Hybrid Data Infrastructure, highlighting the aggregative nature of this kind of infrastructure and the management facility performed though a thin client (a web browser).

'variety' well. It supports a very limited set of data types and needs a common software middleware, dedicated hardware resources and a costly infrastructure management regulated by rigid policies and procedures.

Cloud Computing, instead, provides an elastic usage of resources that are maintained by third-party providers. It is based on the assumption that the management of hardware and middleware can be centralized, while the applications remain in the hands of the consumer. This considerably reduces application maintenance and operational costs. However, as it is a technology based on an agreement between the resource provider and the consumer, it is not suitable to manage the integration of resources deployed and maintained by diverse distributed organizations.

The Hybrid Data Infrastructure (HDI) is a new, more effective solution for managing the new types of scientific dataset. It assumes that several technologies, including Grid, private and public Cloud, can be integrated to provide an elastic access and usage of data and data-management capabilities.

The gCube software system, whose technological development has been coordinated by ISTI-CNR, implements the HDI approach. It was initially conceived to manage distributed computing infrastructures. It has evolved to operate large-scale HDIs enabling a data-management-capability-delivery model in which computing, storage, data and software are made accessible by the infrastructure and are exploited by users using a thin client (namely a web browser), through dedicated on-demand Virtual Research Environments.

gCube operates a large federation of computational and storage resources by relying on a rich and open array of mediator services for interfacing with Grid (eg European Grid Infrastructure), commercial cloud (eg Microsoft Azure, Amazon EC2), and private cloud (eg OpenNebula) infrastructures. Relational databases, geospatial storage systems (eg Geoserver), nosql databases (eg Cassandra, MongoDB), and reliable distributed computing platforms (eg Hadoop) can all be exploited as infrastructural resources. This guarantees a technological solution suitable for the volume, velocity, and variety of the new science patterns.

gCube is much more than a software integration platform; it is also equipped with software frameworks for data management (access and storage, integration, curation, discovery, manipulation, mining, and visualization) and workflow definition and execution. These frameworks offer traditional data management facilities in an innovative way by taking advantage of the plethora of integrated and seamlessly accessible technologies. The supported data types cover a wide spectrum ranging from tabular data – eg observations, statistics, records – to research products – eg diagrams, maps, species distribution models, enhanced publications. Datasets are associated with rich metadata and provenance information which facilitate effective re-use. These software frameworks can be configured to implement different policies ranging from the enforcement of privacy via encryption and secure access control to the promotion of data sharing while guaranteeing provenance and attribution.

The infrastructure enabled by gCube is now exploited to serve scientists operating in different domains, such as biologists generating model-based large-scale predictions of natural occurrences of species and statisticians managing and integrating statistical data.

gCube is a collaborative effort of several research, academic and industrial centres including ISTI-CNR (IT), University of Athens (GR), University of Basel (CH), Engineering Ingegneria Informatica SpA (IT), University of Strathclyde (UK), CERN (CH). Its development has been partially supported by the following European projects: DILIGENT (FP6-2003-IST-2), D4Science (FP7-INFRA-2007-1.2.2), D4Science-II (FP7-INFRA-2008-1.2.2), iMarine (FP7-INFRASTRUCTURES-2011-2), and EUBrazilOpenBio (FP7-ICT-2011-EU-Brazil).

Link:
gCube website:
http://www.gcube-system.org

Please contact:
Pasquale Pagano, ISTI-CNR, Italy
E-mail: pasquale.pagano@isti.cnr.it