Mastering Data-Intensive Collaboration and Decision Making through a Cloud Infrastructure: The Dicode EU project

Collaboration and decision making settings are often associated with huge, ever-increasing amounts of multiple types of data, obtained from diverse sources, which have a low signal-to-noise ratio for addressing the problem at hand. In many cases, the raw information is so overwhelming that stakeholders are often at a loss to know even where to begin to make sense of it. In addition, these data may vary in terms of subjectivity and importance, ranging from individual opinions and estimations to broadly accepted practices and indisputable measurements and scientific results. Their types can be of diverse level as far as human understanding and machine interpretation are concerned.

Nowadays, big volumes of data can be effortlessly added to a database; the problems start when we want to consider and exploit the accumulated data, which may have been collected over a few weeks or months, and meaningfully analyze them towards making a decision. Admittedly, when things get complex, we need to identify, understand and exploit data patterns; we need to aggregate big volumes of data from multiple sources, and then mine it for insights that would never emerge from manual inspection or analysis of any single data source.

Taking the above issues into account, the recently funded Dicode project (http://dicode-project.eu/) aims at facilitating and augmenting collaboration and decision making in data-intensive and cognitively-complex settings. To do so, it will exploit and build on the most prominent high-performance computing paradigms and large data processing technologies - such as cloud computing, MapReduce, Apache Hadoop, Apache Mahout, and column databases – to meaningfully search, analyze and aggregate data existing in diverse, extremely large, and rapidly evolving sources. Services to be developed and integrated in the context of the Dicode project will be released under an open source license.

The Dicode project is timely for the following reasons:

- Cloud computing is making a growing presence in both industry and academia. It is becoming a scalable services delivery and consumption platform for Services Computing (at the same time, services are becoming more and more data intensive). Compared to its predecessors (i.e. grid computing, utility computing), cloud computing is better positioned in terms of economic viability, cost-effectiveness, scalability, reliability, interoperability, and open source implementations.
- There is much advancement in the development of scalable data mining frameworks and technologies (most of them exploiting the cloud computing paradigm), such as MapReduce, Apache Hadoop, and Apache Mahout. Likewise, text mining technologies (such as named entity recognition, named entity disambiguation, relation extraction, and opinion mining) have reached a level in which it is - for the first time - practically feasible to apply semantic technologies to very large data collections, thus allowing capture of an unprecedented amount of information from unstructured texts.
- In parallel, there is much advancement in the development of collaboration and decision making support applications, mainly by exploiting the Web 2.0 features and technologies.
- While helpful in particular problem instances and settings, the above categories of advancements demonstrate a series of limitations and inefficiencies when applied to

data-intensive and cognitively-complex collaboration and decision making support settings.

Building on current advancements, the solution foreseen in the Dicode project will bring together the reasoning capabilities of both the machine and the humans. It can be viewed as an innovative "workbench" incorporating and orchestrating a set of interoperable services (Figure 1) that reduce the data-intensiveness and complexity overload at critical decision points to a manageable level, thus permitting stakeholders to be more productive and concentrate on creative and innovative activities.

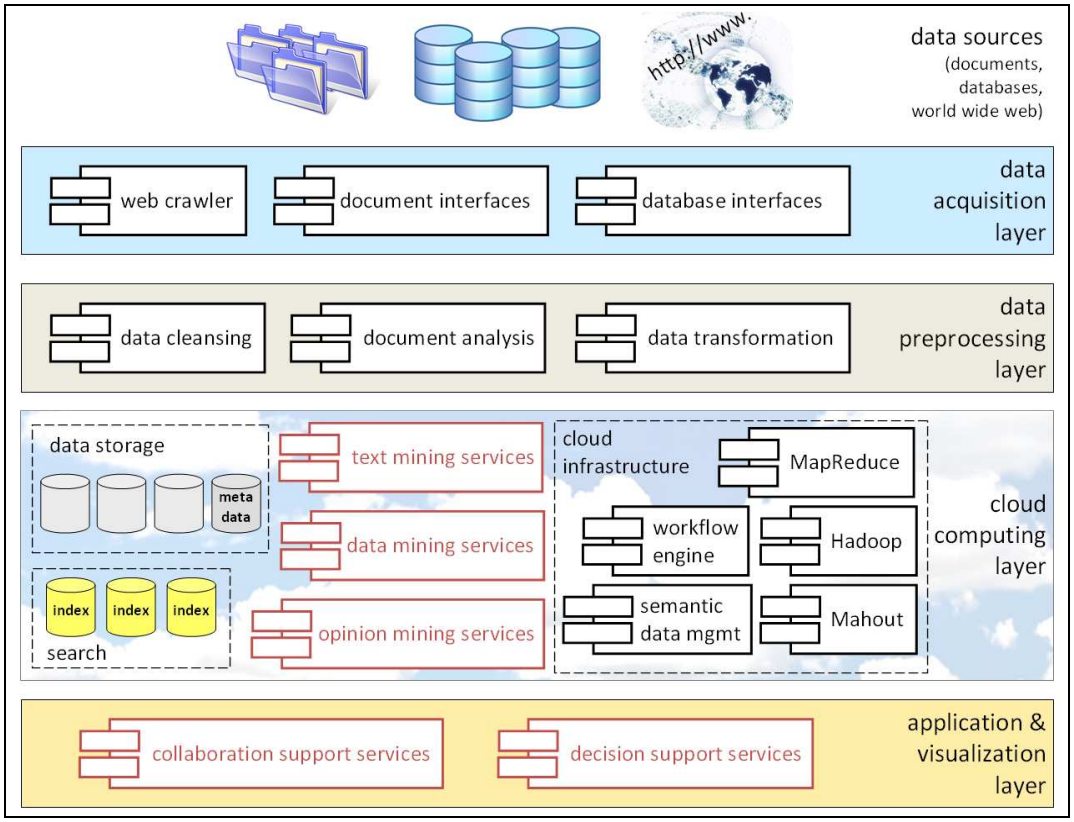

Figure 1: The Dicode Architecture and Suite of Services.

The achievement of the Dicode project's goal will be validated through three use cases. These were chosen to test the transferability of Dicode solutions in different collaboration and decision making settings, associated with diverse types of data and data sources, thus covering the full range of the foreseen solution's features and functionalities. These cases concern:

- *Clinico-Genomic Research Assimilator.* This case will demonstrate how Dicode can support clinico-genomic scientific research in the current post-genomic era. The need to collaboratively explore, evaluate, disseminate and diffuse relative scientific findings and results is more than profound today. Towards this objective, Dicode envisages to plan an integrated clinico-genomic knowledge discovery and decision making use case that targets the identification and validation of predictive clinico-genomic models and biomarkers. The use case is founded on the seamless integration of both heterogeneous clinico-genomic data sources and advanced analytical techniques provided by Dicode.
- *Trial of Rheumatoid Arthritis Treatment*. This case will benefit from Dicode's services to deliver pertinent information to communities of doctors and patients in the domain of Rheumatoid Arthritis (RA). RA treatment trials will be carried out by an academic research establishment on behalf of pharmaceutical company. Each trial will evaluate

the effectiveness of treatment for RA by analysing the condition in wrists (and possibly other joints). Dicode services will be used to enable an affective and collaborative way of working towards decision making by various individuals involved (Radiographers, Radiologists, Clinicians, etc.).

- *Opinion Mining from unstructured Web 2.0 data*. It is paramount today that companies know what is being said about their services or products. With the current tools, finding who and what is being said is literally searching for a needle in the haystack of unstructured information. Through this case, we aim to validate the Dicode services for the automatic analyses of this voluminous amount of unstructured information. Data for this case will be primarily obtained from spidering the Web (blogs, forums, and news). We will also make use of different APIs from various Web 2.0 platforms, such as micro-blogging platforms (Twitter), and social network platforms (Facebook).

The Dicode project is funded by the European Union under FP7 (total cost: 3,510,000 €, funding: 2,600,000 €). It started on September 1st, 2010 and its duration is 36 months.

The partners of the Dicode consortium are: Research Academic Computer Technology Institute (project coordinator), University of Leeds, Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., Universidad Politécnica de Madrid, neofonie Gmbh, Image Analysis Ltd, Biomedical Research Foundation - Academy of Athens, and Publicis Frankfurt GmbH.

Dicode on Twitter: http://twitter.com/DICODE_EU

Dicode on Facebook: http://www.facebook.com/people/Dicode-Eu/100001390513581

Contact persons:

Nikos Karacapilidis
Research Academic Computer Technology Institute
26504 Rio Patras, Greece
Tel: +30 2610 960305
E-mail: karacap@cti.gr

Stefan Rüping
Fraunhofer Institute for Intelligent Analysis and Information Systems
53757 Sankt Augustin, Germany
Tel: +49 2241 143512
E-mail: stefan.rueping@iais.fraunhofer.de

Scott Robinson
neofonie GmbH
Robert-Koch-Platz 4, 10115 Berlin, Germany
Tel: +49 30 246 27 562
E-mail: scott@neofonie.de