Project Synopsis

<Logo> <URL to website>

http://dicode-project.eu/

<Headline>

Exploiting a cloud infrastructure to augment collaboration and decision making in data-intensive and cognitively-complex settings

<Abstract>

The goal of the Dicode project is to facilitate and augment 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 to meaningfully search, analyze and aggregate data existing in diverse, extremely large, and rapidly evolving sources. The foreseen solution can be viewed as an innovative workbench incorporating and orchestrating a set of interoperable services 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 activities. Services to be developed are: (i) scalable data mining services (including services for text mining and opinion mining), (ii) collaboration support services, and (iii) decision making support services. The achievement of the Dicode project's goal will be validated through three use cases addressing clearly established problems. These concern: (i) scientific collaboration supported by integrated large-scale knowledge discovery in clinico-genomic research, (ii) delivering pertinent information from heterogeneous data to communities of doctors and patients in medical treatment decision making, and (iii) capturing tractable, commercially valuable high-level information from unstructured Web 2.0 data for opinion mining.

<Project description> <Image>

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.

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.

Building on current advancements, the solution foreseen in the Dicode project will bring

together the reasoning capabilities of both the machine and the humans (Figure 1). It can be viewed as an innovative "workbench" incorporating and orchestrating a set of interoperable services 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.

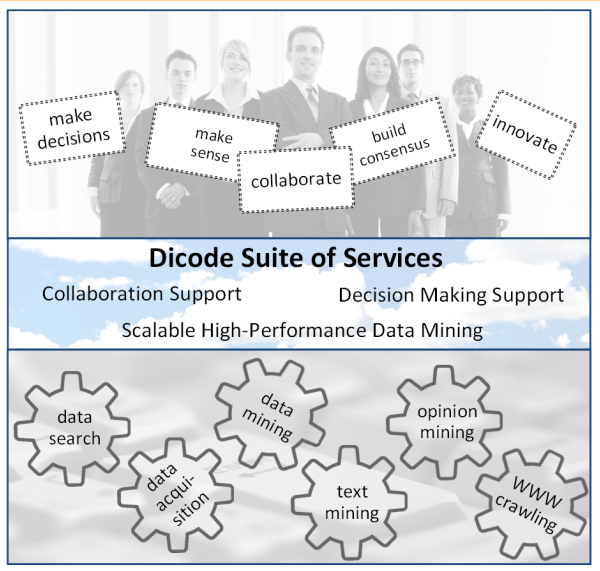

Figure 1: *The Dicode services exploit the cloud computing paradigm and build on the synergy of machine and human reasoning.*

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).