Dicode Annual Report (2010)

http://dicode-project.eu/

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.

Summary of Activities

Major activities that took place during the first four months of the project (September – December 2010) were: (i) user requirements analysis, (ii) state-of-the-art survey, (iii) setup of the basic infrastructure to accommodate the full range of the foreseen Dicode services. These activities are of paramount importance for the fulfillment of the project's objectives. Their successful completion will provide the Dicode consortium with the necessary background knowledge to produce a detailed specification for the overall Dicode architecture and properly start the implementation work scheduled for the next year.

User requirements analysis

Requirements were collected from the three use cases of the project, which 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. Our aim was to understand current practices, focusing on their data intensive decision making and collaborative activities, as well as on the characteristics of data and their usage. A template was designed to help identifying the focus for each use case study and gathering initial sample sets. A rich set of usage scenarios has been elaborated, involving both end users and technical partners. These scenarios helped significantly towards establishing a consensus of the envisioned suite of Dicode services.

State-of-the-art survey

This work area concerns a thorough identification and assessment of issues related to the overall Dicode concept and approach. Issues investigated concern the diversity of solutions that may contribute to the remedy of information overload and cognitive complexity in contemporary business, scientific and social settings. Prominent information and knowledge processing technologies, based on scalable high-performance computing, were assessed. Issues related to collaboration and decision making support were investigated, paying much attention to the end-user perspective. Integration issues, from both a conceptual and a technical point of view, were also explored. The survey was conducted by all technical partners. It is structured according to the three overlapping research directions of the project, namely techniques for scalable high-performance data mining, data mining to make sense of real-world multi-faceted data, and collaboration and decision making support technologies.

Setup of the basic Dicode infrastructure

This work area concerns the setup of the basic infrastructure (i.e. Hadoop, Mahout) on in-house computer clusters for running large scale data mining experiments and testing prototype implementations, as well as the setup of data collections for benchmarking based on textual data (i.e. Apache Software Foundation mailing lists, Wikipedia) and structured data (i.e. Yahoo collection). It also concerns experimentation with the existing Mahout machine learning algorithms and collaboration support platforms. Furthermore, work in this area includes the definition of standards and guidelines for the development of Dicode services, aiming at ensuring interoperability between the services to be developed and reusability of them through diverse scenarios of use. Through this work, partners will reach an agreement on technical requirements, as well as on development technologies and supporting tools to handle issues such as management of source code repository, software building management and bug tracking.

User Involvement, Promotion and Awareness

During the first four months of the project, the Dicode partners representing the user/industry group (i.e. Biomedical Research Foundation - Academy of Athens, Image Analysis Ltd and Publicis Frankfurt GmbH) were extensively involved in the identification and elaboration of user requirements. At the same period, these partners exploited their dissemination channels and promoted the Dicode project.

Up to now, the project’s visibility has been raised through presentations at international conferences, scientific publications and press releases. More specifically, Dicode got coverage (as part of a talk on Mahout) at two major open source conferences: ApacheCon North America 2010, the official user conference of the Apache Software Foundation that took place in Atlanta/GA, as well as Devoxx 2010, the world's largest Java community conference that took place in Antwerp/Belgium. Also, Dicode got coverage at the Codebits event in Lisbon/Portugal. Furthermore, a reference to Dicode now appears on the official blog of the Apache Software Foundation.

Access to the project related publications and press releases is provided through the Dicode web portal (see “Dissemination” page). Finally, Dicode already maintains a profile to widely used social networking sites (Facebook, Twitter, LinkedIn).

Future Work

During the next year, detailed specifications will be produced, operational versions of the Dicode services will be implemented and integrated, innovative work methodologies will be sketched, and feedback from the first validation and assessment of the Dicode outcomes will start to be collected.

An operational integrated suite of Dicode services will be available in 2011 for trials and proof-of-concept purposes. The existence of this suite early on in the project will enable early exploitation of the project’s outcomes, while also augmenting dissemination and ensuring project sustainability.

Being committed to an open source approach, the Dicode project is expected to deliver re-usable results on various levels of intelligent information management.

Further Information

Dicode dissemination efforts → http://dicode-project.eu/index.php?q=news

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

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

Dicode on the ASF's official blog → https://blogs.apache.org/foundation/entry/the_asf_asks_have_you1