DILIGENT Data Challenge on EGEE Infrastructure

Thursday, Dec 06 @ 05:40 EST

The success of DILIGENT's recent data challenge on image feature extraction, executed on the EGEE infrastructure, has enabled one of the world's largest collections of multimedia metadata to be made publicly available for research purposes.

The DILIGENT team used the EGEE computing Grid to process 37 million images from the online Flickr database in just 16 weeks. This computation generated approximately 112 million text and image objects—nearly 5 TB of data—containing more than 150 million extracted features. This is equivalent to an average processing capacity of over 300,000 images per day.

This unique collection will be used by the SAPIR project to develop new large-scale content-based data retrieval and automatic data classification techniques that combine both text and image content, expanding the limits of conventional search engines, which can only search text associated to images and audio-visual content.

The computational load required to generate this massive data collection was outsourced to DILIGENT, and then delegated to the EGEE Pre-Production Service (PPS) Grid infrastructure via the gLite middleware. A total of 44,333 gLite jobs were successfully executed by the EGEE PPS infrastructure resource broker. Each job processed approximately 1000 images.

The data challenge lasted for 116 days, from 16 June to 9 October 2007, and was organized in three different phases. During the initial preparation phase experimental jobs were submitted to some EGEE PPS sites to test the feature extraction application and optimize the number of images to process per day. The next two phases involved actual execution of the data challenge, exploiting ten EGEE PPS sites that contributed their computational resources: University of Athens, Scuola Normale Superiore, ISTI-CNR, LIP, ESA-ESRIN, CERN, CESGA, University of Macedonia, Ben Gurion University, and CYFRONET. Four of these sites are maintained by DILIGENT partners.