DILIGENT Data Challenges

Description

The goal of this data challenge is to execute feature extraction on images. The feature extraction tool that was used is composed by a Java application, some Perl scripts and a C application. The Java code implements a client that contacts the Flickr database (http://www.flickr.com/), downloads a set of users (limited to 5 for interaction) and the images that these users are sharing over the Web. The Perl script and the C application are the core of the feature extraction process - they extract features from the images, create thumbnails and store the results on the CNR site.

More characteristics of the data challenge:

- 1000 jobs submitted per day (through 2 WMSs), although this can be increased/decreased as needed
- Each job processes 1000 images and requires at most 50 Mb of disk space and at least 512 of RAM
- Jobs consume between 20 minutes and 1 hour of CPU time (depending on CPU)
- Sites do not need to install any particular libraries or other software

Schedule

The data challenge total duration was 116 days and was organized in 3 different phases:

- Preparation: From 16 June to 15 July (30 days)
- 1st phase: From 16 July to 29 July (14 days)
- 2nd phase: From 30 July to 9 October (72 days)

During the preparation phase only experimental jobs were submitted since the feature extraction application was being ported and the numbers of images to process per job was being tuned. During this period 3 DILIGENT PPS sites were used. The 1st and 2nd phase correspond to the real execution of the DC. 10 PPS sites were exploited. In the first phase each job contained 500 images to process whereas in the 2nd phase to number was increased to 1000 images.

Results

The following tables present statistics about the execution of the data challenge collected during the 116 days of execution:

Number of jobs

	Submitted	Processed	%
Preparation	7500	5200	69,33
1st phase	7500	5000	66,67
2nd phase	51440	34133	66,35
Total	66440	44333	66,73

Number of input images

	per Job	Expected	Processed	%
Preparation	250	1875000	1300000	69,33
1st phase	500	3750000	2500000	66,67

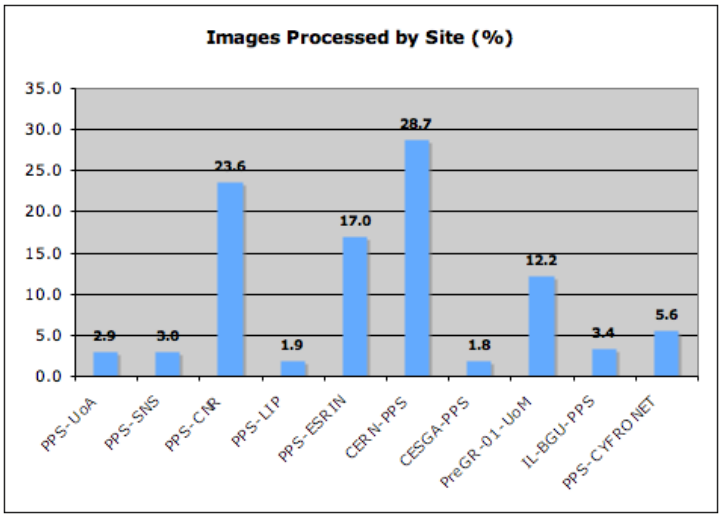
2nd phase	1000	51440000	33408603	64,95
Total		57065000	37208603	65,20

Number of output products

	Generated
Preparation	3900000
1st phase	7500000
2nd phase	100225809
Total	111625809

The ***112 million*** products generated correspond to ***4,55 TB*** of data and contain approximately ***150 million*** features

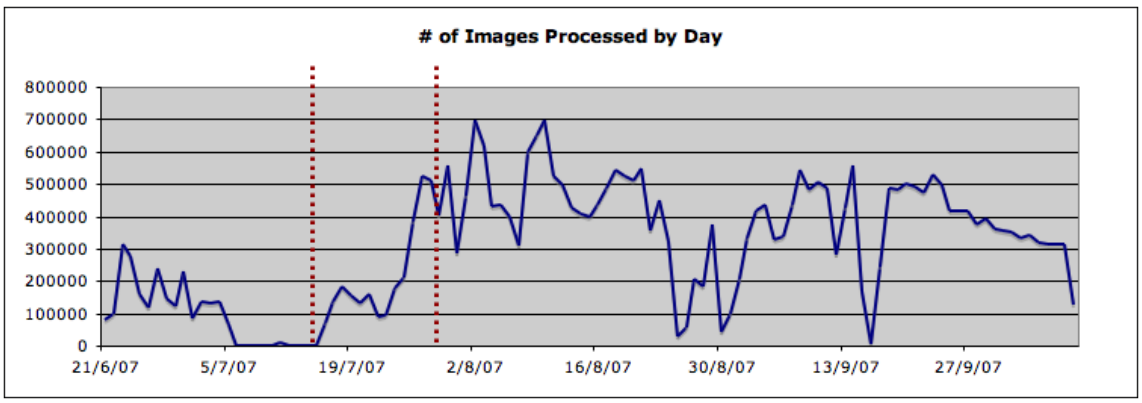
Images processed by site



The feature extraction jobs were being submitted every day. Since the CEs available were changing every day, a number of CEs were not used during some periods of time. Despite this factor, several conclusions can be extracted from the graph above:

1. Only 10 (out of the current 30) PPS sites support the "diligent" VO in this DC
2. Four sites contributed with 80% of the total computation: CERN-PPS, PPS-CNR, PPS-ESRIN, PreGr-01-UoM
3. The DILIGENT PPS sites (PPS-UoA, PPS-SNS and PPS-ESRIN) processed 46,5% of the images

Images processed by day



The red dashed lines in the graph highlight the 3 different phases of the DILIGENT DC: preparation, 1st phase and 2nd phase. During the preparation phase it's visible an initial work load to test the feature extraction tool that then decreased due to the preparation of the real DC jobs. The 1st phase (even if very short) shows a constant increase in the number of images processed. This reflects the strategy of gradually adding new PPS sites. Finally in the 2nd phase (with 1000 image per job) the number of images processed per day slightly increased in the first days and remained around 500 thousand in the rest of the DC. During this long final phase there were 3 short periods where the number of images processed decreased significantly. These low performance periods can be justified by upgrades in the sites (actually they occurred during the end of August / beginning of September when PPS administrators were back from summer holidays!).

Daily Statistics

Daily statistics for each grid node can be be found at:

- http://dlib-services.isti.cnr.it/datachallenge/log_count_dlib.html

-- PedroAndrade - 30 Jul 2007

This topic: DILIGENT > DiligentFlickrDC
Topic revision: r7 - 2007-10-13 - unknown