TaLTaC in ENEAGRID Infrastructure

Silvio Migliori[1], Andrea Quintiliani[1], Daniela Alderuccio[1],
Fiorenzo Ambrosino[1], Antonio Colavincenzo[1], Marialuisa Mongelli[1],
Samuele Pierattini[1], Giovanni Ponti[1]

Sergio Bolasco[2], Francesco Baiocchi[3], Giovanni De Gasperis[4],

[1]ENEA – DTE-ICT – silvio.migliori@enea.it
[2] Sapienza Università di Roma - [3] Staff TaLTac - [4] Dip. DISIM Università dell'Aquila -
info@taltac.it

Abstract

The aim of this joint ENEA-TaLTaC project is to enable the TaLTaC User Community and the Digital Humanists to have remote access to the TaLTaC software through the ENEAGRID Infrastructure. ENEA's research activities on the integration of Language Technologies (Multilingual Text Mining Software and Lexical Resources) in the ENEA distributed digital infrastructure provide a "community Cloud" approach in a digital collaborative environment and on an integrated platform of tools and digital resources, for the sharing of knowledge and analysis of textual corpora in Economic and Social Sciences and e-Humanities. Access to the TaLTac software in Windows and Linux version will exploit the high computational capacity (800 Teraflops) of the e-infrastructure, to which users access as a single virtual supercomputer.

Riassunto

Obiettivo del progetto congiunto ENEA-TaLTaC è consentire alla comunità degli utenti TaLTaC e ai ricercatori nelle Digital Humanities l'accesso remoto al software TaLTaC attraverso l'infrastruttura digitale ENEAGRID. Le attività di ricerca dell'ENEA sull'integrazione delle tecnologie linguistiche (software di Text Mining per testi multilingue e risorse lessicali) in ENEAGRID forniscono un approccio "community Cloud" in un ambiente collaborativo digitale e su una piattaforma integrata di strumenti e risorse digitali, per la condivisione delle conoscenze e l'analisi di corpora testuali in Scienze Economiche e Sociali ed e-Humanities. L'accesso al software TaLTac in versione Windows e Linux sfrutterà l'elevata capacità computazionale (800 Teraflops) dell'infrastruttura di calcolo, a cui gli utenti accedono come ad un unico supercomputer virtuale.

Keywords: Text Mining Software, Cloud Computing, Digital-Humanities, Socio-Economic Sciences, Big Data.

1. Introduction

"TaLTaC in CLOUD" is a joint ENEA-TaLTaC project for the set-up of an ICT portal on the ENEA distributed e-Infrastructure[1] (Ponti *et al.*, 2014), hosting TaLTaC Software (Bolasco *et al.*, 2016, 2017). Users will access TaLTaC software (Windows and Linux versions) in a remote and ubiquitous way, and the computational power (800 Teraflops) of ICT ENEA distributed resources, as a single supercomputer. The aim of this joint ENEA-TaLTaC project is to enable the TaLTaC User Community and Digital Humanists to have remote access to TaLTaC software through ENEAGRID Infrastructure, integrating ICT inside Digital Cultural Research.

ENEAGRID offers a digital collaborative environment and an integrated platform of tools and resources assisting research collaborations, for sharing knowledge and digital resources and for storing textual data. In this virtual environment, TaLTaC software evolves from a stand-alone uniprocessor software toward a multiprocessor design, integrated in an ICT research e-

[1] The ENEAGRID infrastructure is based on several software components which interact with each other to offer an integrated distributed system. The ENEAGRID infrastructure allows access to all these resources as a single virtual system, with an integrated computational availability of about 16000 cores, provided by several multiplatform systems.

infrastructure. Furthermore, it evolves towards implementing ancient language lexical and semantic knowledge and e-resources, facing research needs and implementing solutions also for Digital Humanities communities.

2. TaLTaC Software

The TaLTaC software package, conceived at the beginning of the 2000s, has been progressively developed to date in three major releases: T1 (2001), T2 (2005) and T3 (2016); it is widespread among the text analysis community in Italy and abroad with over 1000 licenses, including two hundred entities between university departments, research institutions and other organizations.

The 2018 release of the software, T3, implemented the following priority objectives: *i*) the processing of big data (around of a billion words), achieving the independence from the dimensions of the text corpora, limited only by hardware resources; *ii*) the automatic extraction on multiple layers of results from text parsing (tokenization): layer zero (text in the original version), layer 1 (recognition of words with automatic corrections of the accents), layer 2 (pre-recognition of most common Named Entities), layer 3 (reconstruction of pre-defined multiwords); *iii)* computing speed, taking advantage of the power of the multi-core processing readily available on current computers (personal or cloud).

Table 1 shows the processing times of three parsing, up to layer 2, for larger corpora on PC (1-core and 8-cores) and on ENEAGRID. Preliminary results on ENEAGRID (1core-CRESCO) show that with increasing corpus size there is an even greater saving of time.

TALTAC was installed in ENEAGRID infrastructure, but the computational capabilities of the HPC system are not yet exploited because the current version of the software does not support multi-core. Therefore, the present ENEAGRID capabilities allow only multi-users access and computation; future versions of the software will be tested for multi-core capabilities to exploit the real power of ENEA ICT High Performance Computing.

Table 1. Preliminary results of processing times of three parsing on PC and on ENEAGRID

				MAC i7 (7th generation)			ENEAGRID
		tokens	size of file	1 core	8 cores	8core /1core	1 core (CRESCO)
		millions	*GB*	*in minutes*		*in %*	*in minutes*
1	"*La Repubblica* " (100 th Artic.)	74	0,41	3,4	1,1	0,33	3,5
2	"*La Repubblica* " (400 th Artic.)	284	1,55	13,0	3,8	0,29	13,2
3	Italian and French Press	535	2,89	37,4	8,8	0,24	41,3
4	Various Press Collection	1.138	6,18	88,2	14,0	0,16	54,7

For the characteristics of the technological architecture of the TaLTaC3 platform, see previous works (Bolasco *et al.* 2016, 2017), that can be summarized here as: a1) *HTML 5 for the GUI* and *jQuery* with its derived Javascript frameworks to encapsulate the GUI user interaction functions for the MAC and Cloud solution; a2) Windows native *DotNET* desktop application; b) *JSON* (*JavaScript Object Notation*): as an inter-module language standard, with a structured and agile format for data exchange in client/server applications; c) *Python / PyPy*: advanced script/compiled programming language, mostly used for textual data analysis and natural language processing at the CORE back end; d) *No-SQL*: high performance key/value data structure storage server *Redis* adopted for vocabularies/linguistic resources persistence; e) *RESTful*: interface standard for data exchange over the HTTP web protocol; f) *MULTI-PROCESSING*: exploiting in the best possible way multi-core hardware, distributing processing power among different CPU cores.

The choice of the Python language allowed to develop a cross-platform computational core running on Windows, Linux, macOS. In particular, the overall system of software processes runs smoothly over a linux-based cloud computing facility, like the ENEAGRID. Furthermore, the Python code compiled through the 64bit *PyPy* just-in-time-compiler allows very efficient macro operations over a large set of data, stored as hash dictionaries, so that the upper limits of performance and capacity is only due to the physical limit of the host machine, in terms of RAM and number of cores and OS kernel scheduler. In our test each node in the ENEAGRID infrastructure hosted a single Redis instance and a number of 24 logic cores, with 16GB of RAM.

3. ENEAGRID Infrastructure

ENEA activities are supported by its ICT infrastructure, providing advanced services as High Performance Computing (HPC), Cloud and Big Data services, communication and collaboration tools. Advanced ICT services are based on ENEA research and development activities in the domains of HPC, of high performance networking and data management, including the integration of large experimental facilities, with a special attention to public services and industrial applications. As far as High Performance Computing is concerned, ENEA manages and develops ENEAGRID, a computing infrastructure distributed over 6 ENEA research centers for a total of about 16000 cores and a peak computing power of 800 Tflops.
HPC clusters are mostly based on conventional Intel Xeon cpu with the addition of some accelerated systems as Intel Xeon/PHI and Nvidia GPU. Storage resources includes RAID systems for a total of 1.8 PB, in SAN/Switched and SRP/Infiniband configuration. Data are made available by distributed and high performances files systems (AFS and GPFS).

ENEA Portici Center has become one of the most important italian HPC center in 2008 with the project CRESCO - Computational RESearch Center for COmplex Systems. CRESCO HPC clusters are used in many of the main ENEA research and developments activities, such as energy, atmosphere and sea modeling, bioinformatics, material science, critical infrastructures analysis, fission and fusion nuclear science and technology, complex systems simulation. CRESCO clusters have provided in 2015 and 2016 more than 40 million core hours each year to ENEA researchers and technologists and to their external partners (external users account for about 30% of the total machine time).

CRESCO6, the new HPC cluster recently installed in Portici in the framework of the 2015 ENEA-CINECA agreement, provides a peak computing power of 700 Tflops and is based on the new 24 cores Intel SkyLake cpu. Its nodes will be connected by the new Intel OmniPath high performance network, providing a 100 Gbps bandwidth.

ENEA ICT department provides also general purpose communication, elaboration and collaboration tools and services as Network management, E-Mail, Video Conferencing and Voip services, Cloud Computing and Storage.

A friendly user access to scientific and technical applications (as Ansys, Comsol, Nastran, Fluent) is provided by dedicated web portals (Virtual laboratories) relying on optimized remote data access tools as NX technology.

4. TaLTaC in ENEAGRID Infrastructure

4.1 Software Installation and Access on ENEA e-Infrastructure

The software TaLTaC is available on Windows and Linux through ENEAGRID via AFS in a geographically distributed file system, which allows remote access to each computing node of the HPC CRESCO systems and Cloud infrastructure from anywhere in the world.
This provides three capabilities: i) data mining, sharing and storage; ii) ICT services necessary for the efficient use of HPC resources, collaborative work, visualization and data analysis; iii) the implementation of software and its settings for future data processing and analysis. Moreover, the availability of the software on the ENEA ICT infrastructure can benefit of the advantages of AFS such as scalability, redundance, backup and so on.

Through the ACL rules it can be possible to manage the accessibility of the software to the community of users in compliance of the license policies that will be put in place. The following two options are provided for TaLTaC running: the first one is to use the applications installed in the windows system and the second one is to use FARO2 – Fast Access to Remote Objects (the general purpose interface for hardware and software capabilities by web access) to directly access the applications installed in the Linux environment and that refer to the data in AFS.

4.1.1. TaLTaC2 (Windows) on Remote Desktop Access

The software TaLTaC2 is available on “Windows Server 2012 R2” by remote desktop access to a virtual machine that can be reached by the ThinLinc general-purpose and intuitive interface. All the users involved in the project activities can access the server but only the person in charge of developing and installing the application can obtain administrator privileges. For this reason, AFS authentication is always required. Every TaLTaC2 user with AFS credentials can access ENEAGRID to run the software and to manage data on AFS own areas via web and from any remote location. In the AFS environment, an assigned disk area with a large memory capacity is defined. This area is mainly used for storage and sharing of large amounts of data (less than 200 MB) (analysis, reports and documents) that come from running the software on a single processor, in serial mode, or for future parallel data mining applications.

4.1.2. TaLTaC3 (Linux) on CRESCO System

On the CRESCO systems, that is accessible from ENEAGRID infrastructure, TaLTaC3 is available on CentOS Linux nodes and then it is possible to leverage the overall computing power dedicated to the activities of TaLTaC and Digital Humanists communities. Every user can start own work session allocating a node with a reserved *Redis* instance and as many computing core as needed.

Performance improvements are obtainable through the parallelization so that a single user can use the full capacity of the assigned node, in terms of number of computing cores. The TaLTaC3 package is automatically started as the user login to the node by a shell script. The opensource Mozilla Firefox web browser makes the user interface in the current beta version. The access to the TaLTaC3 portal use the ThinLinc remote desktop visualization technology that allows an almost transparent remote session on the HPC system, including the graphical user interface, thanks to the built-in features such as load-balancing, accelerated graphics and platform-specific optimisations. Input and output data can be accessed through the ENEAGRID filesystems and therefore easily uploaded and downloaded.

4.2 Case Studies

ENEA distributed infrastructure (and cloud services) enables the management of research process in Economic-Social Sciences and Digital Humanities, providing technology solutions and tools to academic departments and research institutes: building and analyzing collections to generate new intellectual products or cultural patterns, data or research processes, building teaching resources, enabling collaborative working and interdisciplinary knowledge transfer.

4.2.1. TaLTaC User Community

The current (2018) community of TaLTaC over the years aggregated users from the computer laboratories of automatic analysis of texts and text mining, also carried out within the institutional courses of bachelor and magistral degrees, plus Ph.D. students from doctoral degree courses at the universities of Rome "La Sapienza" and "Tor Vergata", of Padua, Modena, Pisa, Naples and Calabria (a total estimate of over 1300 students over the last eight years); furthermore, there is another set of users that subscribed to specific tutorial courses dedicated to TaLTaC (more than 60 courses for a total number of 750 tutorial participants).
A call about the opportunity of using "remotely" the software via the ENEA distributed computing facilities, received the manifestation of interest by 40 departments and other research institutes.

4.2.2. Digital Humanities Community as TaLTaC user

In collaboration with academic experts, ENEA focused on Digital Humanities projects in Text Mining & Analysis in Ancient Writings Systems of the Near East and used TaLTaC2 to perform quantitative linguistic analysis in cuneiform corpora (transliterated into latin alphabet) (Ponti *et al.*, 2017).

Cuneiform was used by a number of cultures in the ancient Near East to write 15 languages over 3,000 years. The cuneiform corpus was estimated to be larger than the corpus of Latin texts but only about 1/10 of the extant cuneiform texts have been read even once in modern times. This huge cuneiform corpus and the restricted number of experts lead to the use of Text Mining and Analysis, clustering algorithms, social network analysis in the TIGRIS Virtual Lab for Digital Assiriology[2], a virtual research environment implemented in ENEA research e-infrastructure. In TIGRIS V-Lab researchers perform basic tasks to extract knowledge from cuneiform corpora. (i.e. dictionaries extraction with word list of toponyms, chrononyms, theonyms, personal names, grammatical and semantic tagging, concordances, corpora annotation, lexicon building, grammar writing, etc.).

5. Conclusions

Researchers and their collaborators will use computational resources in ENEAGRID to perform their work regardless of the location of the specific machine or of the employed hardware/ software platform.
ENEAGRID offers computation and storage resources and services in a ubiquitous and remote way. It integrates a cloud computing environment and exports: a) remote software (i.e. TaLTaC); b) Virtual Labs: thematic areas accessible via web, where researchers can find set of software (and documentation regarding specific research areas); c) remote storage facilities

[2] TIGRIS - Toward Integration of e-tools in GRId Infrastructure for e-aSsyriology http://www.afs.enea.it/project/tigris/indexOpen.php - http://www.laboratorivirtuali.enea.it/it/prime-pagine/ctigris

(with OpenAFS file system). In this virtual environment, TaLTaC software evolves from a uniprocessor software toward a multiprocessor design, integrated in an ICT research e-infrastructure.

This project leads to the TaLTaC evolution from a stand-alone software (allowing Text Mining & Analysis to search for linguistic constructions in textual corpora, showing results in a table or concordance list) to a software "*always and anywhere on*", that also can be accessed, providing an interface where you can visualize results, create interpretative models, collaborate with others, combine different textual representations and storing data, co-developing research practices. Furthermore, this project reflects the shift from the individual-researcher-approach to a collaborative research community-approach, leading to a community-driven software design, tailor-made on specific research community needs and to Community Cloud Computing. This interdisciplinary knowledge transfer enables creating/activating new knowledge from big (cultural and socio-economic) data, both in modern and ancient languages.

References

Bolasco, S., Baiocchi, F., Canzonetti, A., De Gasperis, G. (2016). "TaLTaC3.0, un software multi-lessicale e uni-testuale ad architettura web", in D. Mayaffre, C. Poudat, L. Vanni, V. Magri, P. Follette (eds.), *Proceedings of JADT 2016*, CNRS University Nice Sophia Antipolis, Volume I, pp. 225-235.

Bolasco S., De Gasperis G. (2017). "TaLTaC 3.0 A Web Multilevel Platform for Textual Big Data in the Social Sciences" in C. Lauro, E. Amaturo, M.G. Grassia, B. Aragona, M. Marino. (eds.) *Data Science and Social Research - Epistemology, Methods, Technology and Applications* (series: Studies in Classification, Data Analysis, and Knowledge Organization) Springer Publ., pp. 97-103.

Ponti G., Palombi F., Abate D., Ambrosino F., Aprea G., Bastianelli T., Beone F., Bertini R., Bracco G., Caporicci M., Calosso B., Chinnici M., Colavincenzo A., Cucurullo A., Dangelo P., De Rosa M., De Michele P., Funel A., Furini G., Giammattei D., Giusepponi S., Guadagni R., Guarnieri G., Italiano A., Magagnino S., Mariano A., Mencuccini G., Mercuri C., Migliori S., Ornelli P., Pecoraro S., Perozziello A., Pierattini S., Podda S., Poggi F., Quintiliani A., Rocchi A., Sciò C., Simoni F., Vita A. (2014) "The Role of Medium Size Facilities in the HPC Ecosystem: The Case of the New CRESCO4 Cluster Integrated in the ENEAGRID Infrastructure". In: *Proceedings of the International Conference on High Performance Computing and Simulation, HPCS* (2014), ISBN: 978-1-4799-5160-4.

Ponti G., Alderuccio, D., Mencuccini, G., Rocchi, A., Migliori, S., Bracco, G., Negri Scafa, P. (2017) "Data Mining Tools and GRID Infrastructure for Text Analysis" in "Private and State in the Ancient Near East" Proceedings of the *58th Rencontre Assyriologique Internationale*, Leiden 16-20 July 2012, edited by R. De Boer and J.G. Dercksen, Eisensbrauns Inc. - LCCN 2017032823 (print) | LCCN 2017034599 (ebook) | ISBN 9781575067858 (ePDF) | ISBN 9781575067841.

ENEAGRID http://www.ict.enea.it/it/hpc -

Laboratori Virtuali http://www.ict.enea.it/it/laboratori-virtualixxx/virtual-labs

TIGRIS Virtual Lab http://www.afs.enea.it/project/tigris/indexOpen.php

TaLTaC: www.taltac.it