



Agenzia nazionale per le nuove tecnologie,
l'energia e lo sviluppo economico sostenibile

TaLTaC in ENEAGRID Infrastructure

toward Parallel Text Mining
of Large Text Corpora (from Gigabytes to Terabytes)
In Grid environment

Roma, 15 giugno 2018

Daniela Alderuccio - ENEA DTE-ICT

TaLTaC & Università dell'Aquila



JADT 2018 - International Conference on Statistical Analysis of Textual Data

ENEAGRID Team



CRESCO6 in fase di installazione



ENEAGRID/CRESCO 2018 People

<http://www.eneagrid.enea.it/people/2018EneaGridPeople.html>

Fiorenzo Ambrosino, Giuseppe Aprea, Tiziano Bastianelli, Riccardo Bertini, Irene Bellagamba, Giovanni Bracco, Luigi Bucci, Francesco Buonocore, Marco Caporicci, Michele Caiazza, Beatrice Calosso, Massimo Celino, Marta Chinnici, Antonio Colavincenzo, Aniello Cucurullo, Pietro D'Angelo, Davide De Chiara, Matteo De Rosa, Daniele Di Mattia, Stefano Ferriani, Gianclaudio Ferro, Claudio Ferrelli, Agostino Funel, Dante Giammattei, Marcello Galli, Simone Giusepponi, Roberto Guadagni, Guido Guarnieri, Michele Gusso, Francesco Iannone, Massimo Marano, Angelo Mariano, Giorgio Mencuccini, Silvio Migliori, Marialuisa Mongelli, Patrizia Ornelli, Simonetta Pagnutti, Filippo Palombi, Salvatore Pecoraro, Antonio Perozziello, Samuele Pierattini, Salvatore Podda, Giovanni Ponti, Andrea Quintiliani, Giuseppe Santomauro, Alberto Scalise, Fabio Simoni, Daniele Visparelli.

Between them a special acknowledgement to the people mostly involved in the daily user support and system management:

Fiorenzo Ambrosino, Antonio Colavincenzo, Agostino Funel, Guido Guarnieri, Filippo Palombi, Giovanni Ponti.

ENEA & TaLTaC - Università dell'Aquila



ENEA Italian National Agency for New Technologies,
Energy and Sustainable Economic Development

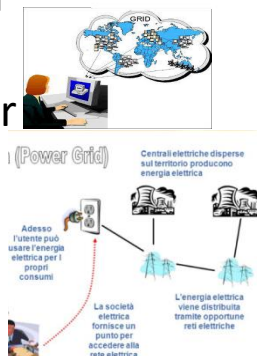
Agenzia nazionale per le nuove tecnologie,
l'energia e lo sviluppo economico sostenibile

ENEA carries out scientific research and technological development by employing a broad range of expertise and advanced plant and instrumental infrastructures located at the 9 Research Centres and 5 Research Laboratories. The headquarters is in Rome. ENEA has a network of regional offices and the Brussels liaison office for participation in European and international research projects. In addition to operating within the Agency's programmes, these infrastructures are also made available to the country's scientific and business communities 2018.

www.enea.it



Supercomputer for
the delivery
of computer power
as a utility
(power grid, water grid)



NEW TECHNOLOGIES: ICT –

Materials Technologies - Radiation Applications - Sustainable Development and Innovation of the Agro-Industrial System;

[ENERGY EFFICIENCY](#)

[RENEWABLE ENERGY SOURCES](#)

[NUCLEAR ENERGY](#)

[CLIMATE AND THE ENVIRONMENT](#)

[SAFETY AND HEALTH](#)

[ELECTRIC SYSTEM RESEARCH](#)

L'infrastruttura HPC di ENEA: **ENEAGRID**
risorse di calcolo e storage distribuite



12 main sites connected by GARR,
the Italian NREN, 6 Research
Centres with HPC/storage resources



TaLTaC

Trattamento automatico
Lessicale e Testuale
del Contenuto
per l'analisi
di un Corpus

TALTAC is the acronym of "Automatic Lexical and Textual Processing for the Analysis of Content" - <http://www.taltac.it/en/taltac1.shtml>

Taltac is a **software application** for the automatic analysis of texts according to the logics of both **Text Analysis** (TA) and **Text Mining** (TM). Such an analysis allows to define a quantitative representation of the phenomenon under study, both at the level of text-units (words) and context-units (words). Consequently, both the **language** and the **contents** of the text can be examined. TALTAC employs both **statistical** and **linguistic** resources.

TaLTaC originates from **research** carried out at the Universities of Salerno and Rome "La Sapienza" during the 1990s under the supervision of Sergio Bolasco, Professor of Statistics at the Department of geo-economic, linguistic, statistical and historical studies for regional analysis of "La Sapienza" University. It is the result of the cooperation of researchers and colleagues of several Italian and French universities.

TALTAC IS widespread among the Text Analysis Community in Italy and abroad with over 1000 licenses, including 200 entities between university departments, research institutions and other organizations.

The TaLTaC software package has been progressively developed to date in three major releases: T1 (2001), T2 (2005) and T3 (2016).

TALTAC3 was developed in collaboration with the DISIM –



Department of Information Engineering, Computer Science and Mathematics at Università degli Studi dell'Aquila



ENEA- TaLTaC Joint Project



Aim: **To enable** Research Communities of:

- ENEA-TaLTaC User
- Digital Humanities
- Socio-Economical Studies

To have **remote access** to the **TaLTaC** software

**Through the ENEAGRID Infrastructure
for Knowledge Sharing and Analysis
of Textual Corpora
in Digital Humanities
and Economic and Social Sciences**

ENEA's research activities on the

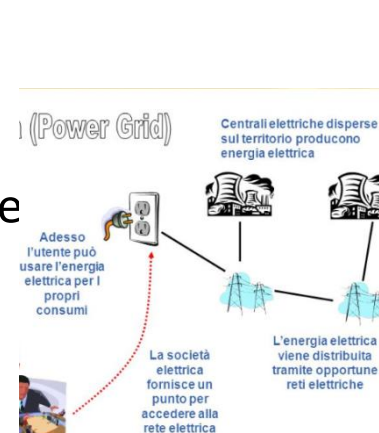
integration of Language Technologies

(Multilingual Text Mining Software and Lexical Resources)

in ENEAGRID - the ENEA distributed digital infrastructure

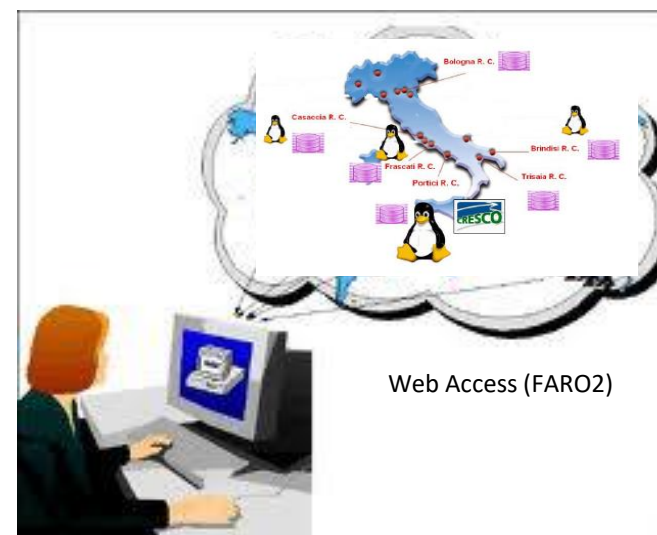
provide a community-cloud approach in a digital collaborative environment

on an integrated platform of tools and digital resources



**Supercomputer for
the delivery
of computer power
as a utility**

(The term *grid computing* originated as a metaphor for making computer power as easy to access as an electric power grid, delivering electricity from producers to consumers)



Why Text Mining in GRID environment?



Parallel and distributed computing emerges as an effective solution for the analysis of large volumes of data, speeding up and scaling up data mining implementation

and for **KNOWLEDGE DISCOVERY in TEXTS**, providing:

- On-line access «always and anywhere ON»
- a simple user interface for users of:
 - TALTAC2 → windows version (web access)
 - TALTAC3 → linux version **for large corpora** on **ENEAGRID** (web access)

(**supercomputers** environment with a “**multi-core**” **architecture**, for large performance benefits)

offering:

- **Data and computing capacity**
- **Virtual Research Environments (virtual labs) & Collaboration tools (network management, video conferencing and voip services, cloud computing, ecc.)**
- **Data storage facilities**
- **Parallel Text/Data Mining & Analysis (data & task parallelism)**

ENEA-TaLTaC - Benefits



1. Researchers will gain **ubiquitous access software and computational resources in ENEAGRID** to perform their work, **regardless:**
 - of the **location** of the specific machine,
 - of the **employed hardware/software platform**.
 - of the **corpus size (from Gigabytes to Terabytes)**.
2. ENEA-TaLTaC project leads to the software evolution from a stand-alone software for Text Mining & Analysis
→ a **software «always and anywhere on»**
3. Co-developing research practices, this project reflects the shifts from the individual-research approach
→ to a **collaborative-research-community approach**
4. This leads → to a **community-driven software design**, tailor-made on specific research community needs and to community cloud computing.

ENEA ÷ TaLTaC partnership



Technological Partner

ENEAGRID digital infrastructure

providing:

- Å computing power and data storage
- Å Saas (Software as a Service)
running Text Mining Tasks in a
parallel & distributed environment
- Å Collaborative research environment

ENEAGRID is open to qualified users.

ENEA Partners access
Collaborative research environment
via credentials (username and password) to
Supercomputers & ICT services:
VPN, Download Software,
Video conferencing, e-Learning, etc).

<http://utict.enea.it/it/servizi-di-base/servizi-di-base/#CredenzAsie>

End-User

Digital Humanities Community

- Å Access TaLTaC software:
 - ï via remote desktop
 - ï via user interfaces to ENEAGRID
- Å Share TaLTaC results :
 - ï via Virtual Laboratories
- Å Storage TaLTaC data in user
environment :
 - ï AFS user folder

ENEA - DTE-ICT



Nell'ambito delle *e-infrastructures* digitali di ricerca,

per lo *storage* e la condivisione dei dati scientifici derivanti da applicazioni in vari settori la Divisione DTE-ICT utilizza **ENEAGRID**: la propria *infrastruttura di supercalcolo scientifico*, <http://www.ict.enea.it/it/hpc>

fortemente integrata e supportata dall'insieme di risorse di rete, servizi di base, servizi *cloud*, strumenti di comunicazione e collaborazione a distanza.

Å Il *cloud-storage* ENEA si appoggia sull'infrastruttura di calcolo ENEA-GRID, che gestisce un sistema di *storage* distribuito e una specifica strategia di backup che ottimizza la conservazione e gestione dei dati.

Å In ENEAGRID è integrato AFS: un ambiente di lavoro comune per l'accesso ai servizi e risorse informatiche, indipendentemente dalla posizione fisica degli utenti

<http://www.eneagrid.enea.it/afs.html>

L'interoperabilità di risorse e strumenti software e hardware garantisce la qualità dei servizi ed un'economia di scala del sistema.

Evolution of HPC resources in Portici:

- 2008 - CRESCO PON 2001-2006 : CRESCO2 -> 25 Tflops (#126 Top500)
- 2013 - CRESCO4 PON 2007-2013: CRESCO4 -> 100 Tflops
- 2018 - CRESCO6 - CINECA Collaboration: CRESCO6 -> 700 Tflops

DTE-ICT: Division of ENEA Energy Technology Dept.

Energy Technology Department (~500 staff)
Information and Communication Technology Division
Services & HPC, (~75 staff) HPC (~20 staff)

•DTE-ICT Services & Activities

- **Basic (Network, AD Authentication, E-mail..)**
- **Administration & Management**
- **Cloud**
- **VMs (~500 on 7 sites)**
- **OwnCloud Storage (ENEABOX)**
Vmware, OpenStack test with GARR
- **Communication tools**
Videoconf (Connect), voip
Phone conf.
- **E-Learning**
- **Web Sites**
- **HPC**
- **Storage, Databases, Big Data**
- **Remote operation of experiments**



12 main ENEA sites connected by GARR, the Italian NREN (National Research & Educational Network)



Text Mining in GRID environment?



TALTAC IN ENEAGRID:

A call about the opportunity of using «remotely» the software via ENEA distributed computing facilities

**received expressions of interest
from TaLTaC User Community:**

40 departments /research institutes answers in 2 days

Area of interests: market, social and opinion research, food, health, political communication, sentiment analysis, etc.

Archiviazione di
grandi quantità di:

- dati,
- e-texts
- ed immagini digitali



Storage



Accesso ai dati



Knowledge Access



Analisi dei dati
e visualizzazione
relazioni interne



Integration of Multilingual
Human Language
Technologies:

- Text Analysis/Mining
- Text Clustering
- Network Analysis and
Visualization



ENEAGRID

Ambiente Digitale interattivo, dall'elevata potenza computazionale (800 Teraflops), a cui l'utente accede come ad un unico supercomputer virtuale.

Ospita :

- **Risorse di calcolo** eterogenee
- **Laboratori Virtuali** – ambienti virtuali collaborativi accessibili «anytime & anywhere» - (Clima, Fluidodinamica, Fissione Nucleare, Scienze dei Materiali, Grafica 3d, e-Humanities, etc.)
<http://utict.enea.it/it/laboratori-virtualixxx/virtual-labs>
- **Software scientifici** tra cui anche software HLT (Human Language Technologies):
 À
 À **TaLTaC2 & TaLTaC3**
- **Collaboration tools** (network management, video conferencing and voip services, cloud computing, ecc.)

ENEAGRID: risorse di calcolo e storage distribuite



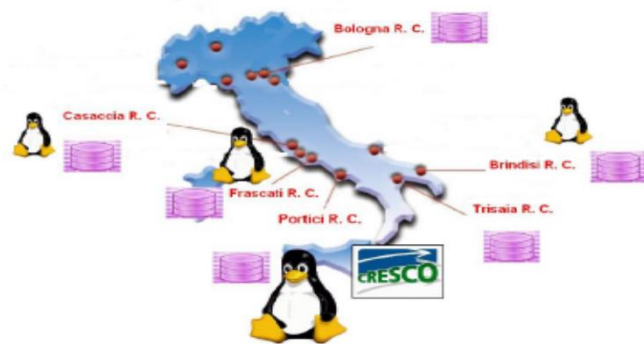
L'infrastruttura HPC di ENEA: **ENEAGRID**
risorse di calcolo e storage distribuite



- <http://www.eneagrid.enea.it>
- <http://www.cresco.enea.it>

Computational Research Center on Complex Systems

12 main sites connected by GARR,
the Italian NREN, 6 Research
Centres with HPC/storage resources



Services & Activities

- HPC CRESCO clusters
- ENEAGRID Infrastructure
- Storage, Databases, Big Data
- Scientific Software
- 3D/Graphic tools & resources
- Remote operation of experiments & lab data preservation
- Virtual Labs
- Application area: material science for energy, fission & fusion nuclear technology, climate and environmental science, complex systems,...

Integrazione servizi & infrastrutture



I servizi ICT

- Servizi di base
- Gestionale
- Servizi cloud
 - Macchine virtuali
 - Storage (ENEABOX)
- Strumenti di comunicazione
 - videoconferenze, voip
 - teleconferenze
- E-Learning
- Teleoperazione esperimenti

Workshop CO.B.RA. Roma 26 gennaio 2016

Research area 2017	Usage (%)
Nuclear energy	33.6
Climate/Env.	25.3
Mat. Sciences	22.7
CFD/Combust.	10.8
Bio Tech.	3.1
Complex Systems	1.7
Aerospace	0.6
Other domains	2.33

2017 - CRESCO clusters provided more than 34 millions core*hours/year = ~ 3880 years of 1 PC 24h/7days to ENEA researcher and external partners



Inauguration: 30th May 2018

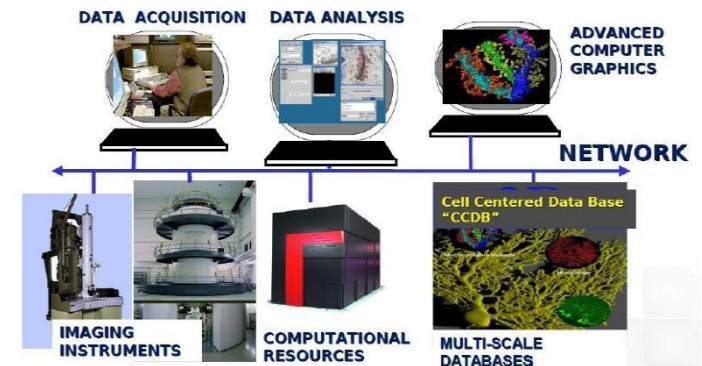
The integration of CRESCO6
in ENEAGRID is currently in progress

~10000 cores

~700 Tflops peak computing power

(+ CRESCO3 + CRESCO = 800 Tflops)

La piattaforma ICT



Workshop CO.B.RA. Roma 26 gennaio 2016

ENEAGRID ĩ Inaugurazione CRESCO6



CRESCO Project ĩ Centro computazionale di RicErca sui Sistemi Complessi

<http://www.cresco.enea.it/>

CRESCO6: risorsa di calcolo di rilievo nel panorama della ricerca scientifica italiana per l'esecuzione di codici paralleli



Frutto di una partnership tra ENEA e CINECA finalizzata alla fornitura a EUROFUSION (il Consorzio Europeo per lo Sviluppo dell'Energia da Fusione), di

servizi di supercalcolo e storage dati

- CRESCO6 offre una potenza computazionale di ~ 700 Teraflops e andrà ad affiancarsi a:
- CRESCO4 (100 Teraflops)
- e CRESCO5 (25 Teraflops),

già installati ed operanti presso il medesimo Centro.

ENEA

High-Profile Technologies

On a persistent
Digital Infrastructure

Evolution of HPC resources in Portici:

- 2008 - CRESCO PON 2001-2006 : CRESCO2 -> 25 Tflops (#126 Top500)
- 2013 - CRESCO4 PON 2007-2013: CRESCO4 -> 100 Tflops
- 2018 - CRESCO6 - CINECA Collaboration: CRESCO6 -> 700 Tflops

Supporto alle attività di Ricerca e Sviluppo nei settori istituzionali dell'Agencia ENEA, così come nelle collaborazioni con enti nazionali ed internazionali e con il sistema produttivo italiano

ENEAGRID for Research Communities



External user affiliation

UnISA	2
Univ Madrid	1
UnLaSapienza	5
ISPRA	1
Consorzio RFX	1
Unicamerino	1
Univ Virginia	1
Vinča Institute Belgrade	1
CREATE Consortium	1
UnIRC	1
UnFI	1
INM-Germany(Julich)	1
UnINA	1

From e-Science to other Research Communities :

- À TaLTaC Users
- À e-Humanities
- Heritage Science
- Big Data
- Web Mining
- OSINT
- ...

ENEAGRID DATA & BIGDATA Sources

The availability of significant amount of computed results c application of the Big Data, Deep learning techniques

Examples of the data available in ENEAGRID/CRESCO

- Climate/energy related forecasts (~100TB)
- **Air pollution models and air quality evaluation (~200 TB, 7x10⁶ files)** collab. with
- **Web crawling (~15 TB, 340x10⁶ web pages)::** presentation in the Workshop
- Nuclear fusion experiments
- HPC Cluster monitoring system (MySQL, 9 GB recording) ZABBIX
- Bioinformatics
-

ENEA G.Bracco, Workshop Big Open Data Analysis, Roma Sede 2/2/

ENEA HPC participation in H2020 projects

Topics of interest and opportunities:

(2016 -> ETP4HPC member)

- HPC infrastructures & technologies, applications (material science), distributed data management, I/O performances, collaboration and remote access tools...
- Many different systems for testing and benchmarking

Running projects:

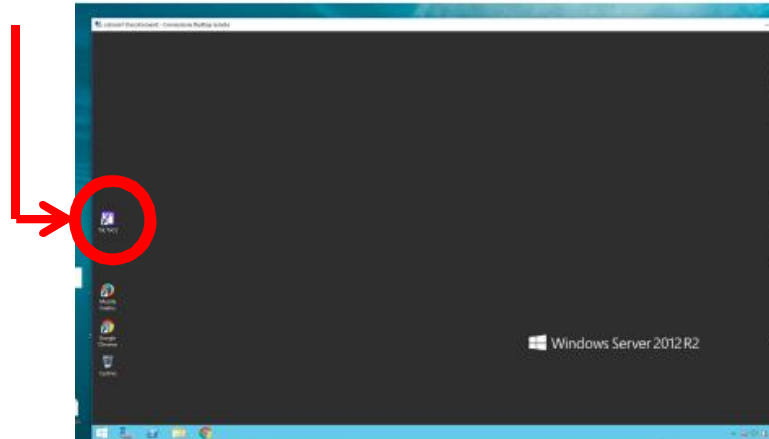
- **H2020 Centre of Excellence: EoCoE** - <http://www.eocoe.eu/>
 - The European Energy oriented Centre of Excellence for computing applications. EoCoE delivers services to exploit the HPC computing power to accelerate the transition to carbon-free energy technologies in industries and society.
- **Support tools** for several H2020 projects: SEADATACLOUD, NEXTOWER, M4F, INSPYRE, GEMMA... and to EERA-JPNM (Joint Programme on Nuclear Materials): e.g. Report Management

Friendly & Ubiquitous Access to TaLTaC



**TaLTaC2 (windows)
on Remote Desktop Access**

TaLTac2



**TaLTaC3 (Linux)
on CRESCO System - www.cresco.enea.it**



TaLTaC Access in ENEAGRID Infrastructure



TaLTaC2 (windows) on Remote Desktop Access

TaLTac software is available on «Windows Server 2012 R2» by remote desktop access to a virtual machine, reached by the ThinLinc general-purpose and intuitive interface.

All users involved in the project activities can access the server.

AFS authentication is always required . (AFS is a distributed network file system facilitating stored server file access between AFS client machines located in different areas)

TaLTaC3 (Linux) on Cresco System (accessible from ENEAGRID infrastructure)

TaLTac software is available on CentOS Linux nodes.

Input and Output data can be accessed through the ENEAGRID filesystems and therefore easily uploaded and downloaded.

Every user can start own work session allocating a node with a reserved Redis instance and as many computing cores 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 open source Mozilla Firefox web browser makes the user interface in the current beta version.

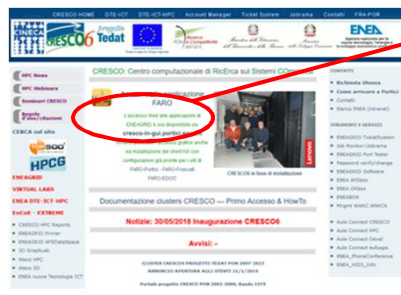
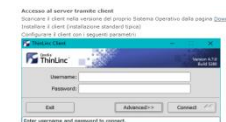
The access to the TaLTaC3 portal use the Thinc Linc remote desktop visualization technology, that allows an almost transparent remote session on the HPC system, including the graphical user interface, thanks to the built-in features such as load-balancing, accelerated graphics and platform-specific optimizations.

How To Access in ENEAGRID Infrastructure



www.cresco.enea.it

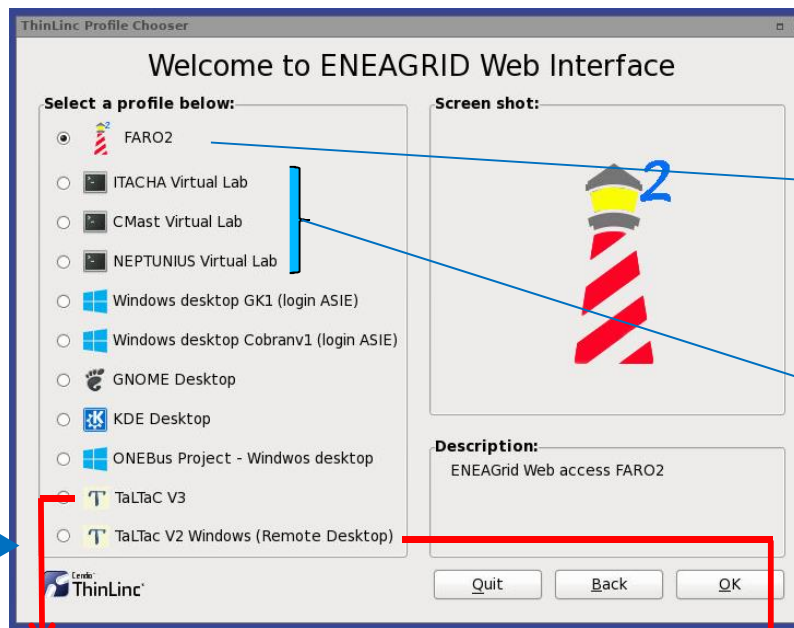
Click on cresco-in-gui.portici.enea.it



ENEAGRID credentials:
ENEA.IT/Username
Password



login form integrated with
ENEAGRID
<https://cresco-in-gui.portici.enea.it:300/main/>



Before accessing www.cresco.enea.it
Download ThinLinc <https://www.cendio.com/thinlinc/features>
ThinLinc is a Cendio software making computing resources available to those who need it, when they need it, for a more efficient use of hardware ;enabling users to move easily between machines , while still being presented with the same desktop . Since all the horsepower resides in the server hall, users no longer need their own expensive hardware to perform even the most *resource-intensive tasks*

FARO2 - Fast Access to Remote Objects
– Web Access Interface –
Remote access to: CRESCO software
ENEAGRID v-labs

Virtual Labs

Software

TaLTac3 (Linux) on Cresco System

TaLTac software is available on CentOS Linux nodes.
Input and Output data can be accessed through the ENEAGRID filesystems and therefore easily uploaded and downloaded.

TaLTac2 (windows) on Remote Desktop Access

TaLTac software is available on «Windows Server 2012 R2» by remote desktop access to a virtual machine, reached by the ThinLinc general-purpose and intuitive interface.

Friendly -Ubiquitous Access to TaLTaC2 (windows version) in ENEAGRID



TaLTaC2 (Windows) Remote Desktop Access

via ThinLinc

<https://www.cendio.com/thinlinc/features>

Web Access

Click on **cresco-in-gui.portici.enea.it**



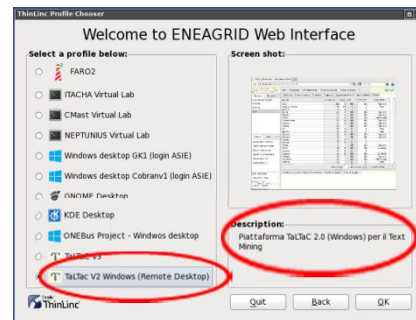
A Cendio software making computing resources available to those who need it, when they need it, for a more efficient use of hardware; enabling users to move easily between machines, while still being presented with the same desktop. Since all the horsepower resides in the server hall, users no longer need their own expensive hardware to perform even the most resource-intensive tasks



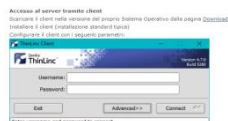
login form integrated with ENEAGRID
<https://cresco-in-gui.portici.enea.it:300/main/>



welcome screen after login
through a browser



Before accessing www.cresco.enea.it
Download ThinLinc



ENEAGRID credentials:
ENEA.IT/Username Password

Windows – Menu - START

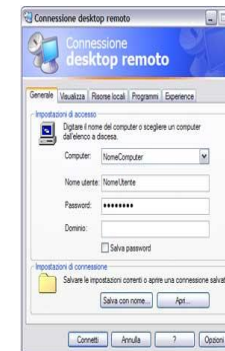
VPN

-- >ACCESSORI WINDOWS

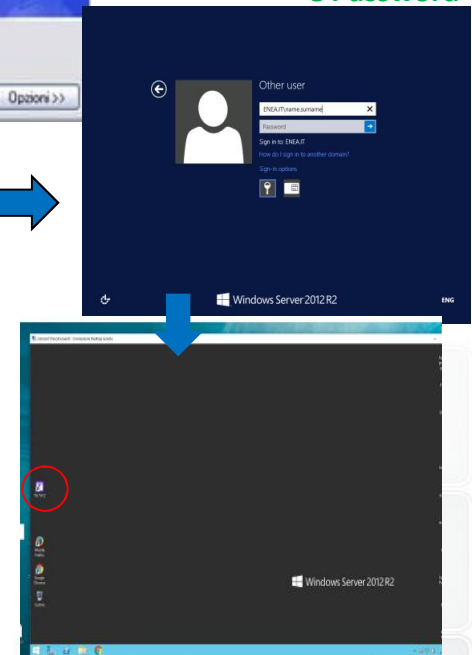
→ Connessione Desktop remoto



ENEASIE credentials:
ENEA.IT/name.surname
e Password



TaLTaC2



Schermata interfaccia di accesso a remote desktop

Daniela Alderuccio – TaLTaC in ENEAGRID - JADT 2018 – Roma, 15 giugno 2018

Friendly & Ubiquitous Access to TaLTaC (windows version)

TaLTaC2 (windows)

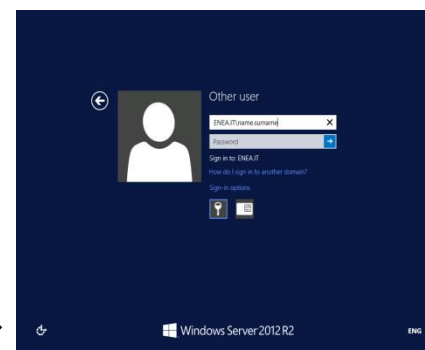
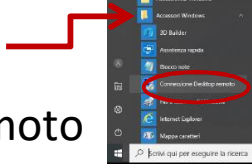
on Remote Desktop Access

VPN

Menu START

-- >ACCESSORI WINDOWS

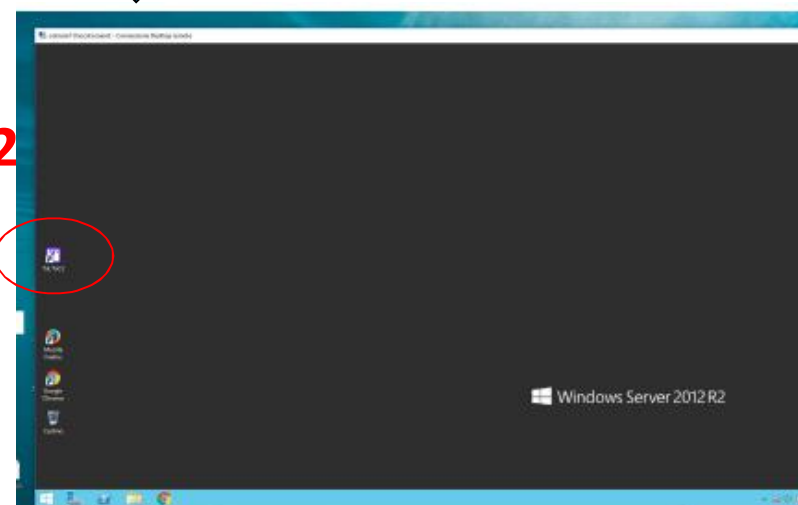
→Connessione Desktop remoto



ENEA ASIE credentials:
ENEA.IT/name.surname
Password



TaLTac2



TaLTaC2 (windows) on Remote Desktop Access

TaLTac software is available on «Windows Server 2012 R2» by remote desktop access to a virtual machine, reached by the ThinLinc general-purpose and intuitive interface. All users involved in the project activities can access the server. AFS authentication is always required . (AFS is a distributed network file system facilitating stored server file access between AFS client machines located in different areas)

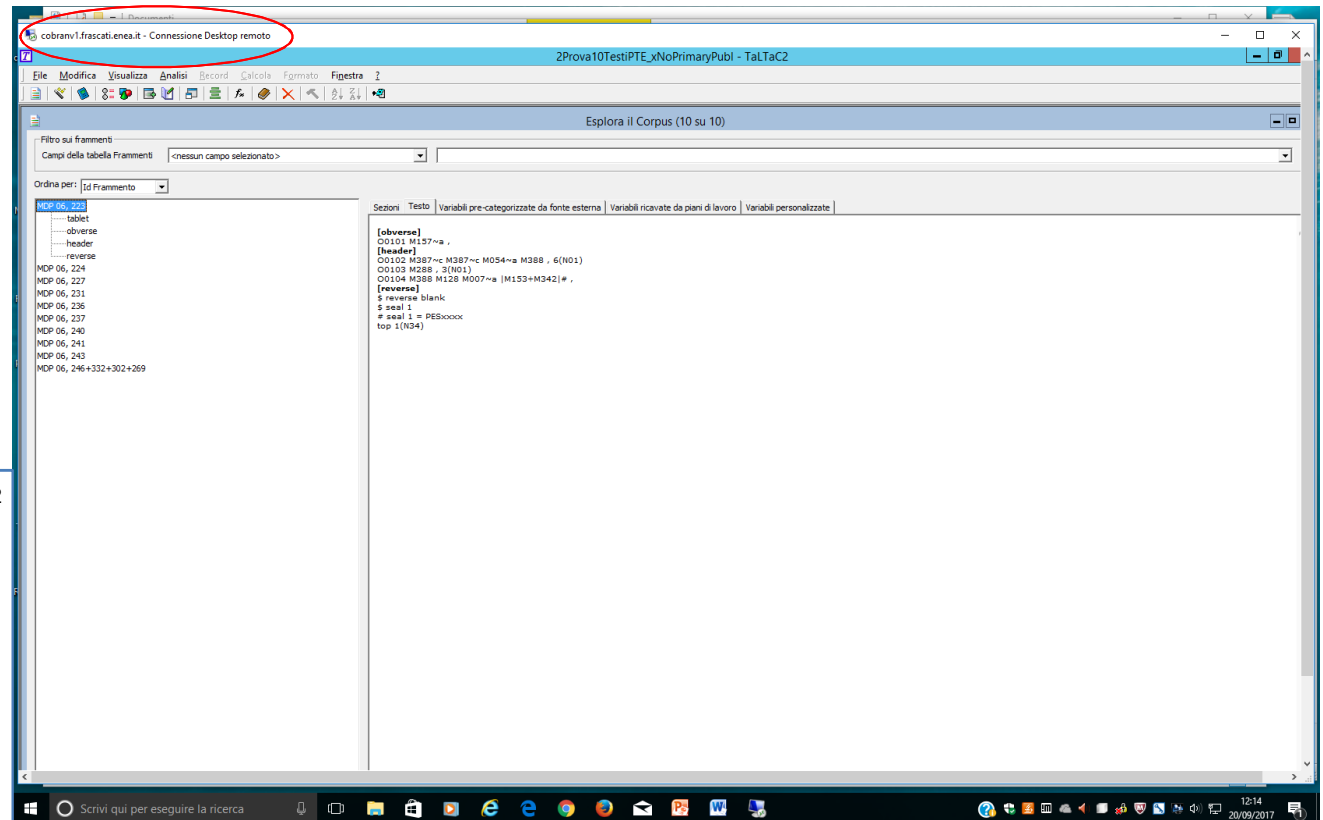
Source: CDLI

MDP 06, 223
 Museo/Institution: MDP 06, 223 (6th-4th V., 1909)
 Collection: Lower
 Museum: Sb 04832
 CDLI no: P008022
 Provenience: Susa



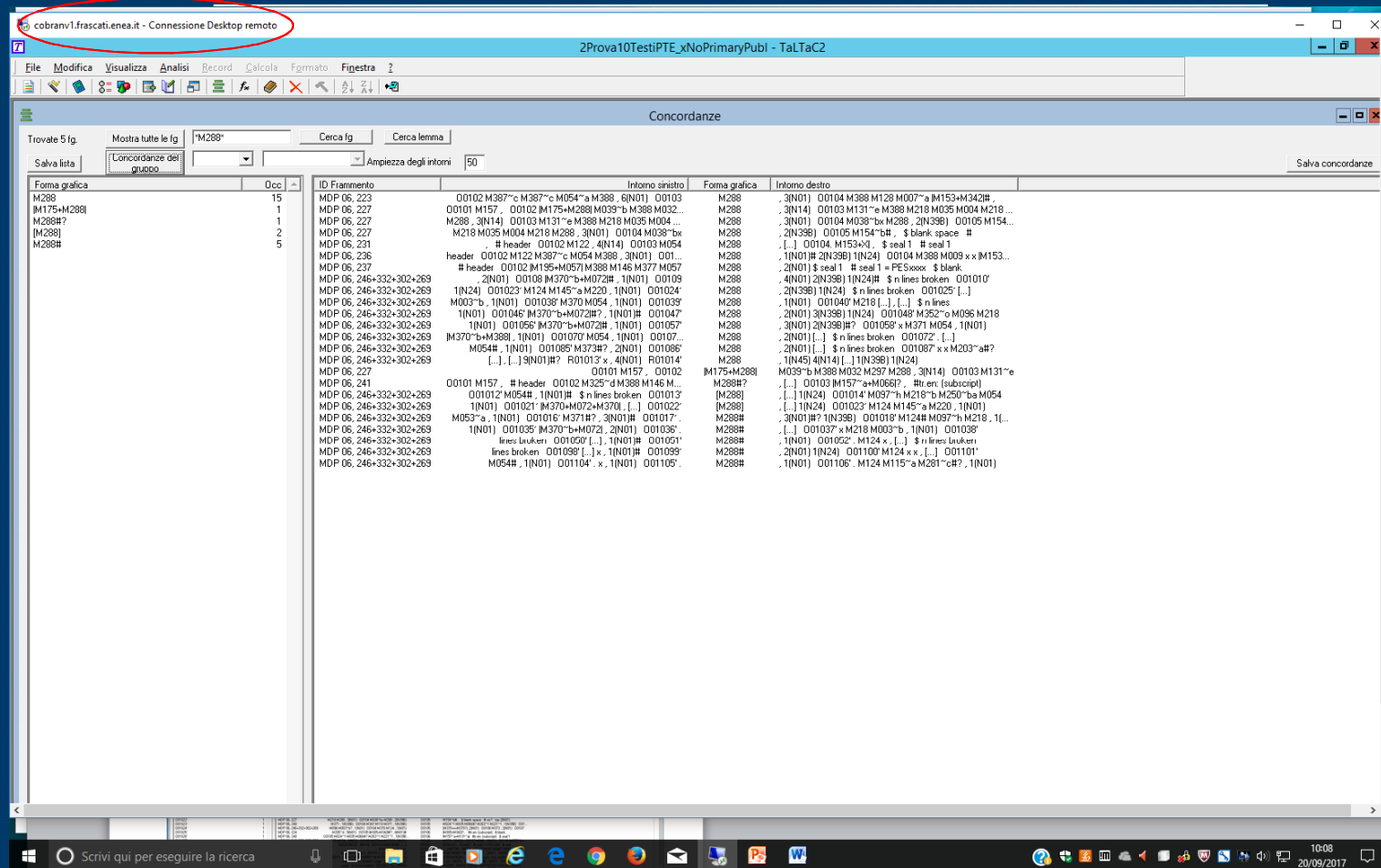
00101 M157
 2 M387, M387, M054 M388 6,
 3 M388 3N,
 4 M388 M128 M007~a |M153+M342|#
 top 1N4

Test TALTaC2 on remote desktop i PROTOELAMITE (3400-2500 BC)



****MDP 06, 223 *MuseumNo=Sb 04832 *CDLI no=P008022
 *Content=Workers and cereals: M388, M387, M288
 *Period=Proto-Elamite(ca.3100-2900BC)
 *Provenience=Susa (mod. Shush) *Genre=Administrative
 *atf=lang qpc
 ++++tablet
 ++++obverse
 00101 M157~a ,
 ++++header
 00102 M387~c M387~c M054~a M388, 6(N01)
 00103 M288, 3(N01)
 00104 M388 M128 M007~a |M153+M342|# ,
 ++++reverse
 \$ reverse blank
 \$ seal 1
 # seal 1 = PESxxxx
 top 1(N34)

MDP 06, 223 i Museum No. Sb 04832 i CDLI no P008022



Concordances of «M288» - «barley» in Protoelamite

Test TALTAC2 (windows version) on remote desktop i PROTOELAMITE (3400-2500 BC)

TALTAC2 (windows) ì



Å Corpus Rep10milaArtANSI.txt size: ~**40 MB**.

Tempo di esecuzione del parsing **2m:10sec**

ì Numero forme del vocabolario: 170774

ì Occorrenze totali del corpus: 6372638

Å Corpus 92_ANSIcert.txt size: ~**150 MB**.

Tempo di esecuzione del parsing **8m:15sec**

ì Numero forme del vocabolario: 339525

ì Occorrenze totali del corpus: 24983109

Remote Desktop allows Multi-Users access without software installation
for the end user

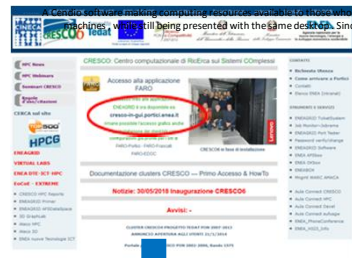
Friendly -Ubiquitous Access to multicore TaLTaC (Linux version) in ENEAGRID



TaLTaC3 (Linux) on Cresco System <http://www.cresco.enea.it/>

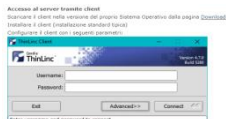
via ThinLinc <https://www.cendio.com/thinlinc/features>

Click on cresco-in-gui.portici.enea.it

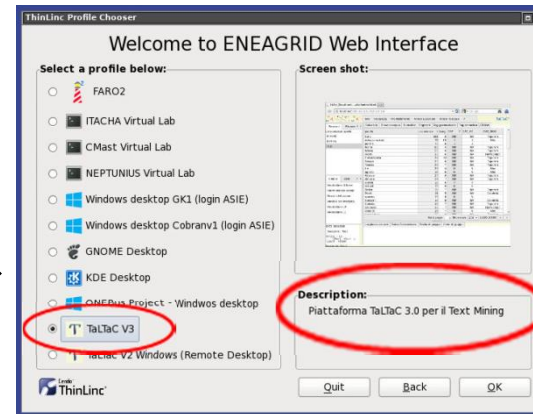


login form integrated with ENEAGRID
<https://cresco-in-gui.portici.enea.it:300/main/>

Before accessing www.cresco.enea.it
Download ThinLinc



welcome screen after login
through a browser



parola	occorrenze	Lung	CAT	Y	CAT_AC	CAT_SEM
Italia	151	6	NM		NM	Toponimi
extracomunitari	72	15	J		J	Altro
perché	70	6	J		J	
Roma	60	4	NM		NM	Toponimi
Milano	58	6	NM		NM	Toponimi
Pietro	39	6	NM		NM	Nome propri
Caltanissetta	34	13	NM		NM	Toponimi
Europa	30	6	NM		NM	Toponimi
Tunisia	29	7	NM		NM	Toponimi
line	29	4	N		N	Altro
agosto	27	6	N		N	Altro
Kosovo	24	6	NM		NM	Toponimi
Albania	24	7	NM		NM	Toponimi
scorla	23	6	n		n	
miliardi	23	8	N		N	
Sicilia	22	7	NM		NM	Toponimi
Prodi	21	5	NM		NM	Celebrità
numero	19	6	N		N	
Gaspari	17	8	NM		NM	Celebrità
Catania	16	7	NM		NM	Toponimi
Secondo	16	7	NM		NM	Nome propri
veneti	15	7	N		N	Altro

Daniela Alderuccio – TaLTaC in ENEAGRID - JADT 2018 – Roma, 15 giugno 2018

TaLTaC Access in ENEAGRID Infrastructure



TaLTaC3 (Linux) on Cresco System (accessible from ENEAGRID infrastructure)

TaLTac software is available on CentOS Linux nodes.

Input and Output data can be accessed through the ENEAGRID filesystems and therefore easily uploaded and downloaded.

Every user can start own work session allocating a node with a reserved Redis instance and as many computing cores 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 open source Mozilla Firefox web browser makes the user interface in the current beta version.

The access to the TaLTaC3 portal use the Thinc Linc 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 optimizations.

TaLTaC3 (Linux) Access on Cresco System <http://www.cresco.enea.it/>

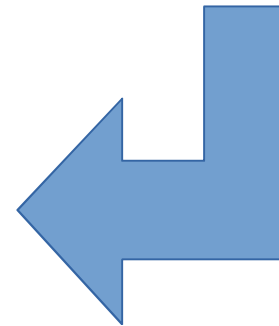


Username:

Password:

Version 4.9.0 (build 5775) on cresco-in-gui.portici.enea.it

Copyright © Cendio AB 2018



http://localhost...atic/index.html

localhost:5000/static/index.html

File Visualizza Pre-trattamento Analisi Lessicale Analisi Testuale ?

Salva tab. Browser corpus Normalizz Segmenti Tag grammaticale Tag semantico Cfr liste

Sessioni Risorse

Lista sessioni aperte

(nuova)

(indice)

Test

Corpus Liste

Vocabolario di base

Frammenti del corpus

Sezioni del corpus

Variabili con modalità

Vocabolario L0

Vocabolario L1

DATI SESSIONE

Sessione: Test

Strato : L1

((VocL, t0cc) =

(11073, 71788)

Sessione-Dati: Test.L1

parola	occorrenze	Lung	CAT	Y	CAT_AC	CAT_SEM
Italia	151	6	NM		NM	Toponimi
extracomunitari	72	15	J		J	Altro
perché	70	6	J		J	
Roma	60	4	NM		NM	Toponimi
Milano	58	6	NM		NM	Toponimi
Pietro	39	6	NM		NM	Nomi propri
Caltanissetta	34	13	NM		NM	Toponimi
Europa	30	6	NM		NM	Toponimi
Tunisia	29	7	NM		NM	Toponimi
lire	29	4	N		N	Altro
agosto	27	6	N		N	Altro
Kosovo	24	6	NM		NM	Toponimi
Albania	24	7	NM		NM	Toponimi
scuola	23	6	n		n	
miliardi	23	8	N		N	
Sicilia	22	7	NM		NM	Toponimi
Prodi	21	5	NM		NM	Celebrità
numero	19	6	N		N	
Gasparri	17	8	NM		NM	Celebrità
Catania	16	7	NM		NM	Toponimi
Secondo	16	7	NM		NM	Nomi propri
venerdi	15	7	N		N	Altro

Go to page: 1 Show rows: 100 1-100 of 538

Larghezza co-testi Salva Concordanze Scelta di gruppo Conc. di gruppo

TALTAC3 in ENEAGRID

Sessioni di lavoro

Area primaria di risultato

Strutture lessicali

Dati testometrici di sessione

Area di dettaglio

http://localhost...atic/index.html

localhost:5000/static/index.html

File Visualizza Pre-trattamento Analisi Lessicale Analisi Testuale ?

Salva tab. Browser corpus Normalizz Segmenti Tag grammaticale Tag semantico Cfr liste

parola	occorrenze	Lung	CAT	Y	CAT_AC	CAT_SEM
Italia	151	6	NM		NM	Toponimi
extracomunitari	72	15	J		J	Altro
perché	70	6	J		J	
Roma	60	4	NM		NM	Toponimi
Milano	58	6	NM		NM	Toponimi
Pietro	39	6	NM		NM	Nomi propri
Caltanissetta	34	13	NM		NM	Toponimi
Europa	30	6	NM		NM	Toponimi
Tunisia	29	7	NM		NM	Toponimi
lire	29	4	N		N	Altro
	27	6	N		N	Altro
Kosovo	24	6	NM		NM	Toponimi
Albania	24	7	NM		NM	Toponimi
scuola	23	6	n		n	
miliardi	23	8	N		N	
Sicilia	22	7	NM		NM	Toponimi
Prodi	21	5	NM		NM	Celebrità
numero	19	6	N		N	
Gasparri	17	8	NM		NM	Celebrità
Catania	16	7	NM		NM	Toponimi
Secondo	16	7	NM		NM	Nomi propri
venerdì	15	7	N		N	Altro

Go to page: 1 Show rows: 100 1-100 of 538

Larghezza co-testi Salva Concordeanze Scelta di gruppo Conc. di gruppo

DATI SESSIONE

Sessione: Test

Strato : L1
(|Voc|, t0cc) =
(11073, 71788)

Multi-core ON/OFF



Dal menu File è possibile attivare e disattivare la modalità MULTICORE, che permette di sfruttare al massimo le capacità di calcolo disponibili sui nodi di ENEAGRID.

Quando è abilitata la suddetta opzione, le operazioni di PARSING (TOKENIZZAZIONE) vengono distribuite su tutti i processori (fisici e logici) disponibili sul nodo, partizionando rispettivamente il corpus in analisi.

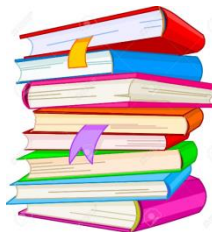
Al termine del calcolo, vengono assemblati i risultati dei vari cores per la costruzione del vocabolario di base.

Le ulteriori strutture lessicali, sono automaticamente costruite tramite procedure di normalizzazione ed etichettatura, grazie alle risorse linguistiche presenti in TaLTaC.

CORPUS & DATABASE Reconstruction?



Large Corpora



ENEAGRID

distributed computing environment

1 Large Corpus

1 User /many Nodes

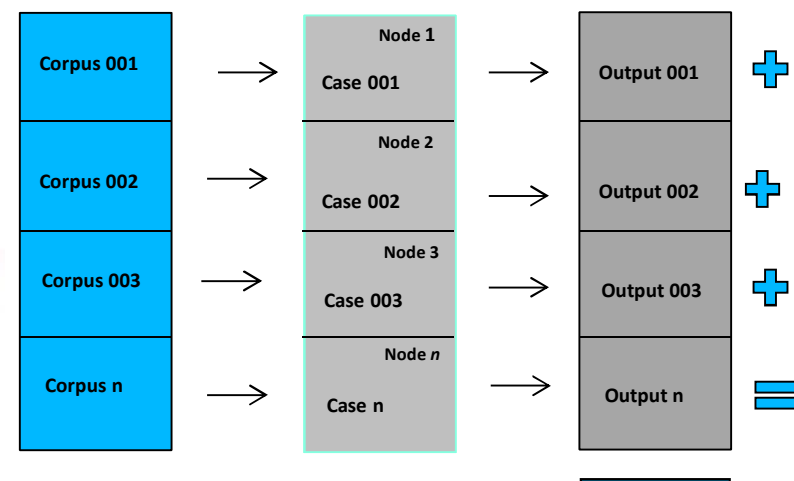
(Simultaneously from 1 to n)

1 Node $\approx$ 1 PC

1 software installation



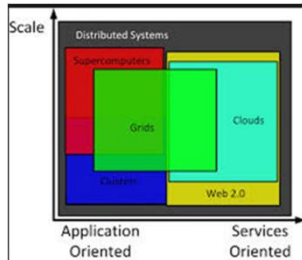
Parallelism



CORPUS &
DATABASE
Reconstruction

1 User Folder

Future Steps



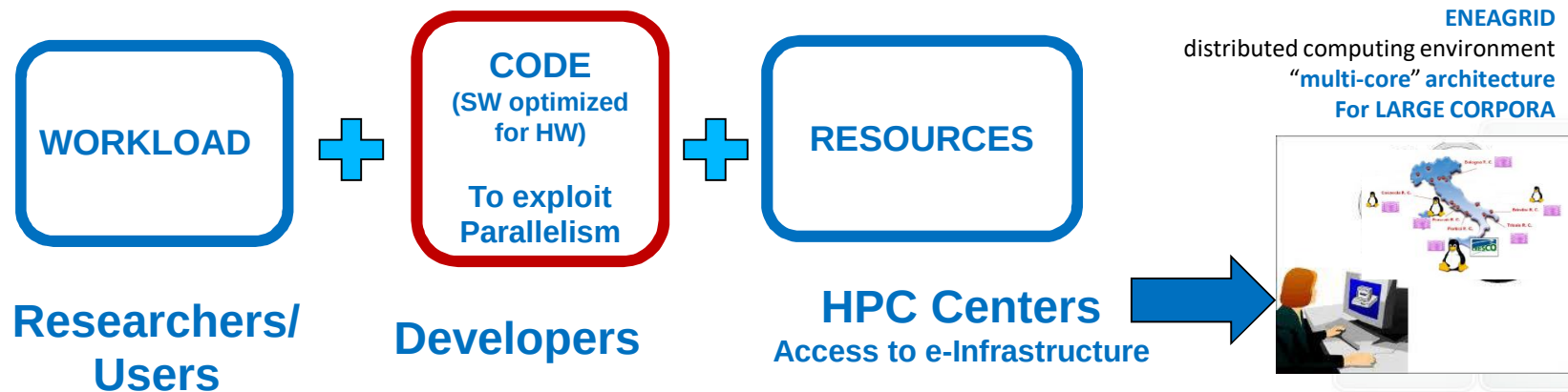
Interest in supercomputing is now worldwide, and growing in many new markets (~50% of Top500 computers are in industry).

Source: Jack Dongarra, «Inaugurazione CRESCO6» 30 maggio 2018

The real power of ENEAGRID will be fully exploited when software will be in MULTI-CORE version
CODE (SW optimized for HW)
To exploit Parallelism

To fully exploit Parallelism in Text Mining in GRID e-Infrastructure
an ADAPTATION OF SOFTWARE APPLICATION IS REQUIRED,
to maximize every form of parallelism within a supercomputer
and use thousands cores simultaneously to solve one large problem

Co-design new CODE Design



Co-Evolution



2 Kilobytes: Typewritten page

17 Kilobytes: The size of an average Web Page

1 Megabyte: 2012 average web page – a small novel

400 Gigabytes: 20,000 trees made into paper and printed

1 Terabyte 250 million pages printed both sides

Kilobyte - 1,024 bytes

Megabyte - 1000 Kilobytes

Gigabyte - 1000 megabytes

Terabyte - 1000 gigabytes

1 Megabyte: 3-1/2 inch diskette

20 Megabytes: Typical hard drive in the first desktop PCs

750 Megabytes: 1 CD

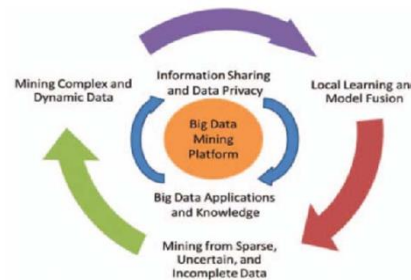
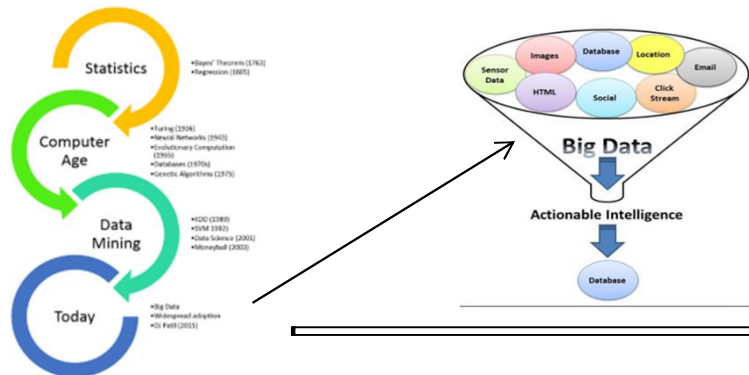
7 Gigabytes: 1 DVD

500 Gigabytes: 100 DVD Movies

1 Terabyte: about 1400 CD-Rom



Data Mining



CRESCO6

ENEAGRID

Conclusions



1. Researchers will gain **ubiquitous access software and computational resources in ENEAGRID** to perform their work, **regardless**:
 - of the **location** of the specific machine,
 - of the **employed hardware/software platform**
 - of the **corpus size (from Gigabytes to Terabytes)**.
2. ENEA-TaLTaC project leads to the software evolution from a stand-alone software for Text Mining & Analysis to **a software «always and anywhere on» integrated in ENEAGRID Infrastructure**
3. **Through the experience gained with TaLTaC we are confident that our technology can be employed for other software applications performing textual analyses.**
4. **ENEAGRID approach** is an **Open Collaboration** addressing **Research Communities interested in Parallel Text Mining of Massive Volume of Text Data** (TB & PB-sized corpora): i.e. TaLTaC User Community and other research communities from various disciplines and tasks: e-Humanities, Heritage Science, Social Sciences, Big Data, Web Mining, Open Source Intelligence, etc.

enabling co-developed research practices, adopting a **collaborative-research-community approach with an interdisciplinary knowledge transfer,**

Creating and activating new Knowledge from Big (cultural and socio-economic) Data

*“Computational approaches can establish meaningful relationships
between a given signal in large-scale textual corpora and verifiable historical moments,*

*but the **understanding** of the **implications** of these findings for people cannot be automated,
and will always be **the realm of the humanities and social sciences**,
and never that of machines.*

***Data-driven approaches can complement the traditional method
in detecting trends of continuity and change in large-scale textual corpora.***

***Data Science enables cross disciplinary research
exploring the interplay between Social Science, Humanities, and large-scale data-driven AI.”***

Nello Cristofanini – Professor of AI University of Bristol

Artificial Intelligence, Machine Learning, Media Content Analysis, Big Data, Epistemological and Ethical Implications of Data-Driven Science and Society
“Leggere 180 milioni di parole” da Tuttoscienze – La Stampa 16 maggio 2018 -large-scale analysis of historical newspapers, modern news, social media content and images

Contacts



DTE-ICT Authors:

Silvio Migliori, Andrea Quintiliani,
Daniela Alderuccio, Fiorenzo
Ambrosino, Maria Luisa Mongelli,
Samuele Pierattini, Giovanni Ponti

TaLTaC Authors:

Sergio Bolasco – uniroma1
Francesco Baiocchi - ISTAT
Giovanni De Gasperis - univaq

For collaboration & information
please contact:

silvio.migliori@enea.it

For any information, please contact:

sergio.bolasco@uniroma1.it

Thanks for your attention!