Examples of Big Data analytics in ENEA:
data sources and information extraction strategies

Ing. Giovanni Ponti, PhD
ENEA – DTE-ICT-HPC

giovanni.ponti@enea.it

DISRUPTIVE DATA 2017
5 Maggio, 2017, Via Santa Maria in Gradi, 4, 01100 Viterbo VT

Outline

- Big Data: intro
- The ENEA context
  - Research fields & application domanis
  - ENEAGRID environment and CRESCO HPC clusters
  - Big Data sources
    - Quantitative statistics on data
    - Data handling and analysis problems
- Data Analytics and Deep Learning tools

Big Data: definitions

Three definitions:

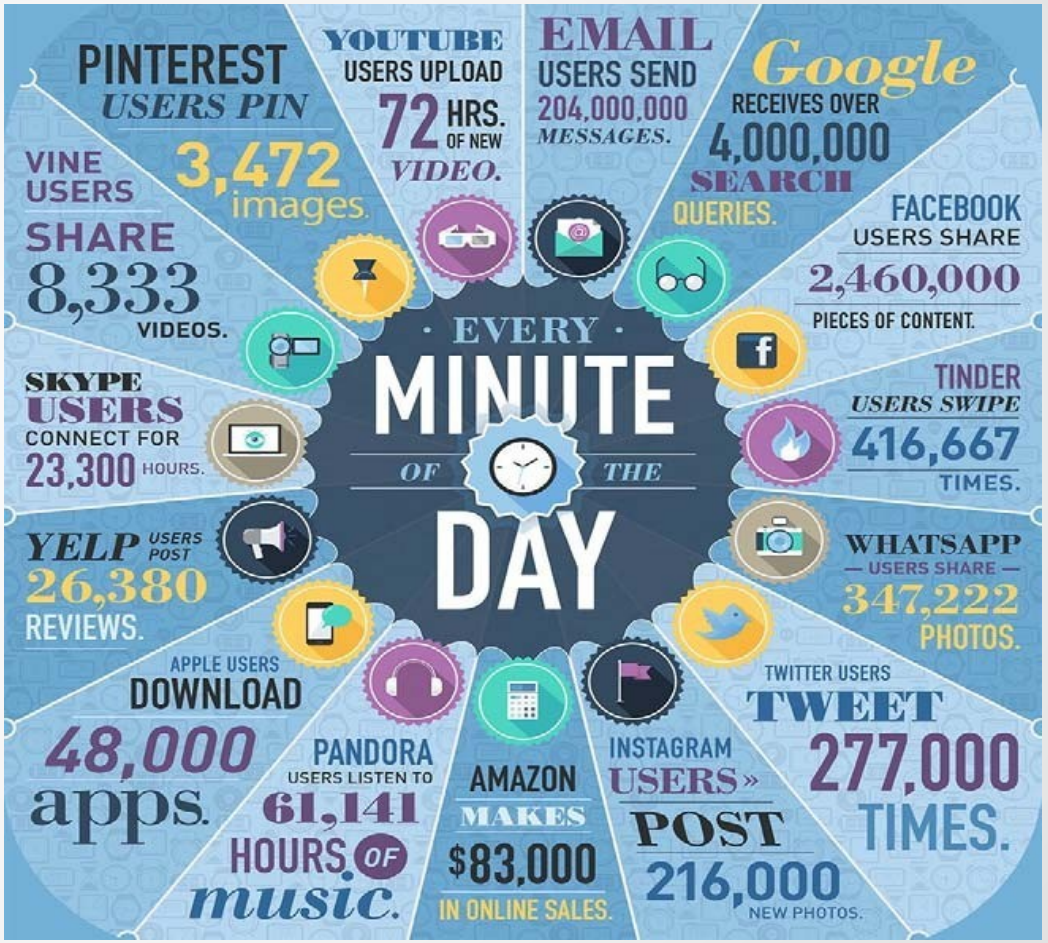
"Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a *tolerable elapsed time* for its user population." - Teradata Magazine article, 2011

"Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze."
- The McKinsey Global Institute, 2012

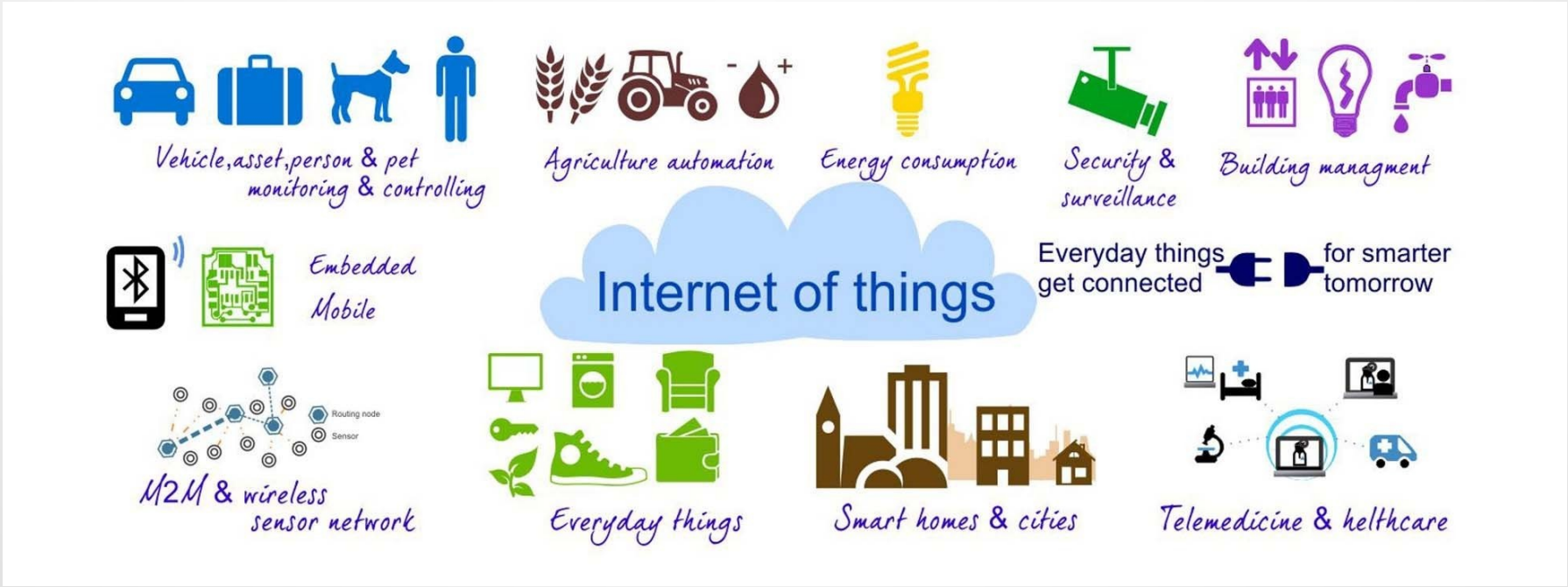
"Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools."
- Wikipedia, 2017

Big Data: some numbers

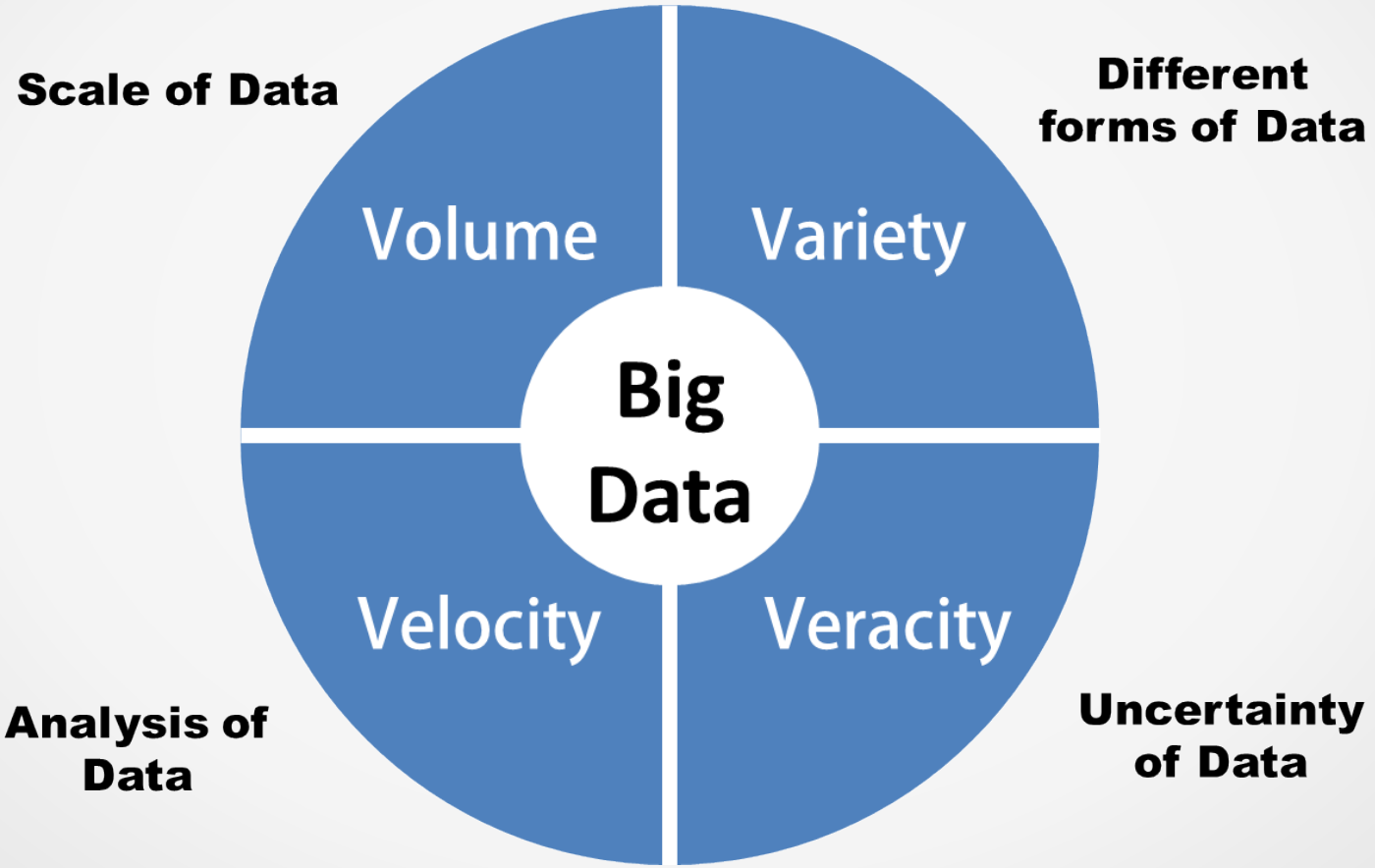
- **How many data in the world?**
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006 (1EB = 10^18B)
  - 4.5 Zettabytes, 2013 (1ZB = 10^21B)
  - 44 Zettabytes by 2020
- **How much is a zettabyte?**
  - A stack of 1TB hard disks that is 25,400 km high
- **How many data in a day?**
  - 2.5 Exabytes
  - 8 TB, Twitter
  - 50 TB, Facebook
- **90% of world's data:**
  - generated over last two years!



Big Data: device proliferation & IoT



Big Data: the 4 “V”s

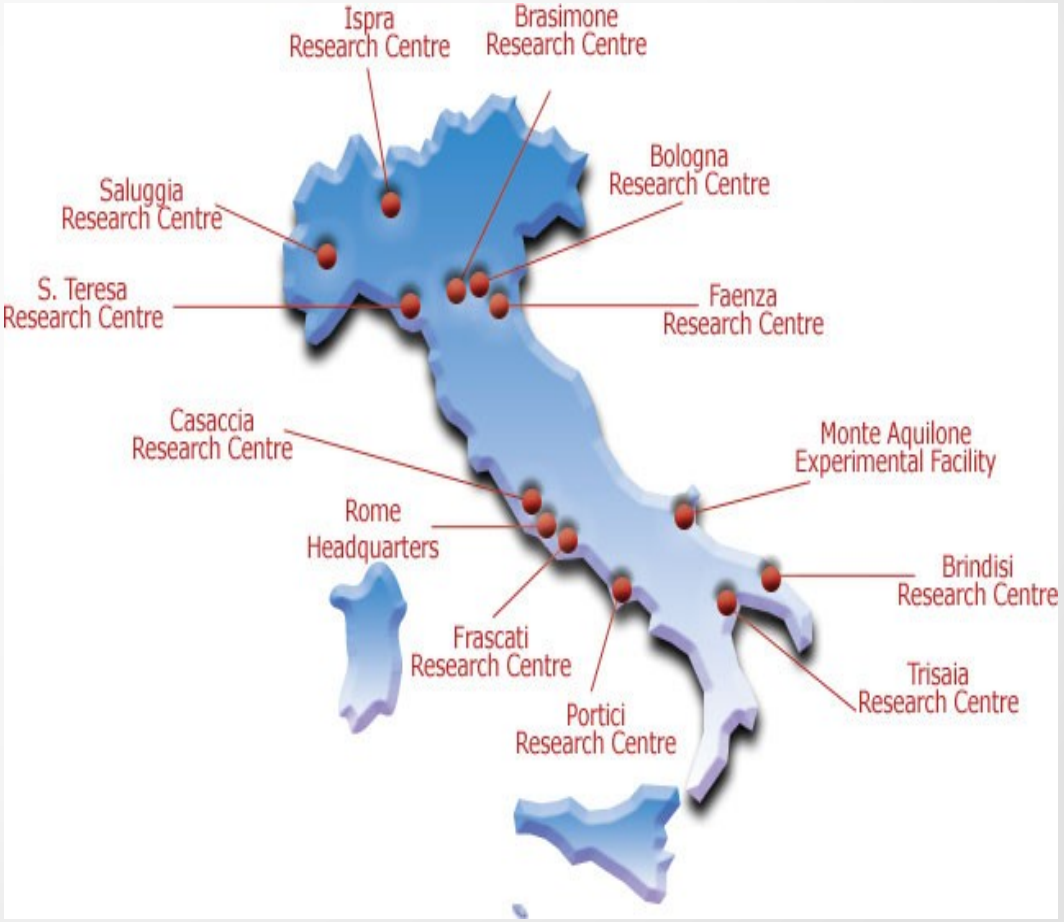


The scenario: ENEA

ENEA is the "*Italian National Agency for New Technologies, Energy and Sustainable Economic Development*"

Research & Development:

- Energy Efficiency
- Renewable Energy Sources
- Nuclear Energy
- Climate and the Environment
- Safety and Health
- New Technologies
- Electrical System Research

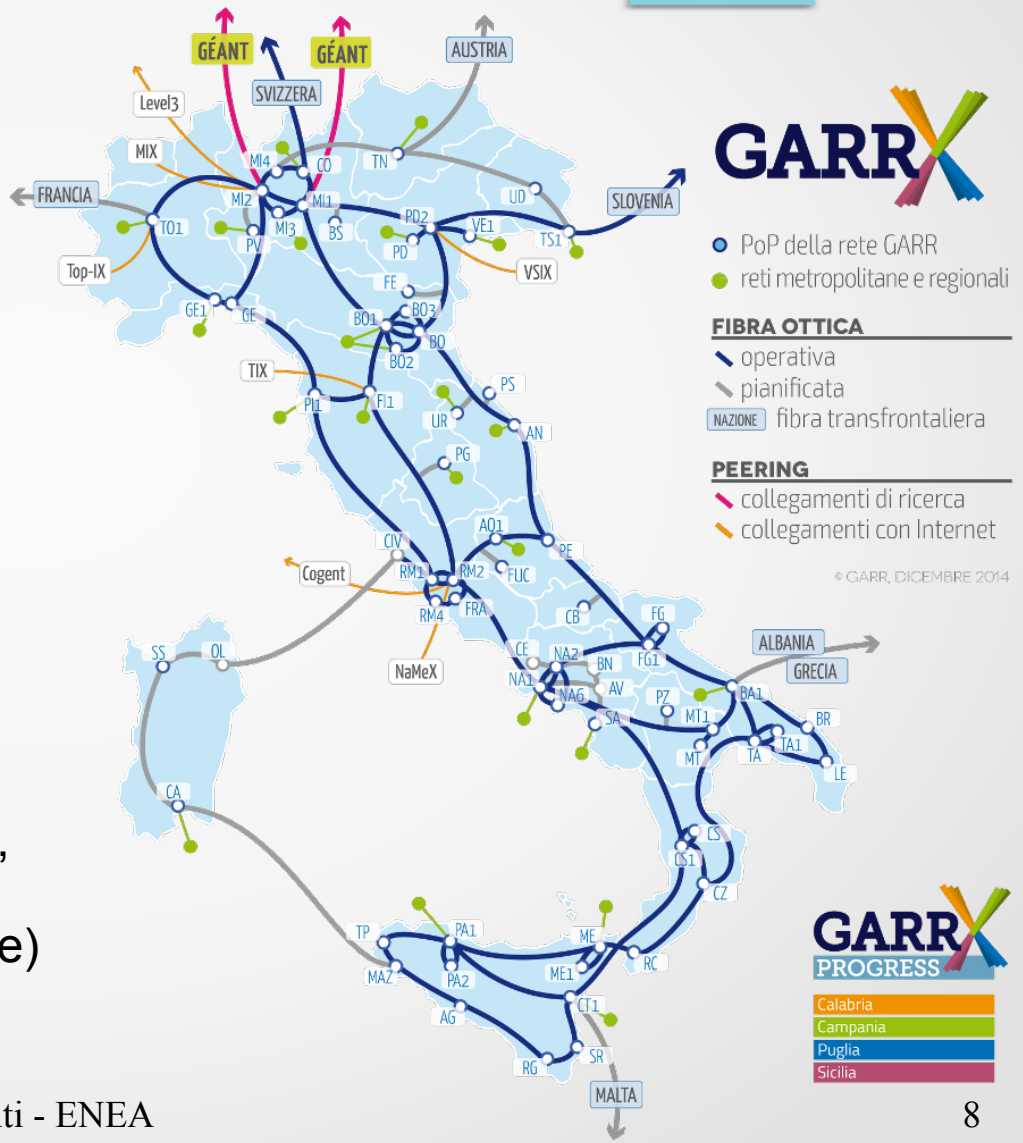


ENEAGRID & CRESCO HPC Clusters

ENEAGRID
Computation & Storage
ENEA distributed resources
interconnected via
GARR network

CRESCO HPC Clusters:

- More than 9000 cores
- Computing nodes:
  - Linux x86_64
  - Special systems (GPU, PHI)
- Storage resources:
  - AFS (distributed)
  - GPFS (parallel high-speed)
- Cloud computing facilities (Openstack, VMWare)
- 6 CED in ENEA (Portici is the main site)



Big Data sources in ENEA

ENEA researcher activities produce every day large amount of data.

Data are stored and elaborated exploiting ENEAGRID computing resources.

Big Data sources in ENEA (main):

- CRESCO monitoring systems
- ENEAGRID user accounting
- Climate forecasts
- Air pollution models
- Web crawling
- Nuclear fusion

CRESCO monitoring systems (1) ZABBIX

Zabbix monitoring tool

- Tool to monitor and track complex large-scale datacenters
  - Computing nodes
  - Network HW
  - Storage HW and services
- Open source www.zabbix.com

CRESCO monitoring systems (2) ZABBIX

ENEAGRID/CRESCO Data

- For each monitored component, data are recorded at *different resolution levels* (form 1 up to 30 min)
- Historical data are stored for 365 days with resolution of 1h – aggregate values of min, mean, and max)

- Zabbix database (MySQL):
  - 88 tables
  - More than 111ML of tuples
  - ~8.5GB
  - Loop recording

ENEAGRID users accounting data (1)

LSF job scheduler

- Workload management platform by IBM for distributed HPC environments
- Allows to define queues and resource types to submit user jobs
- Set of intelligent, policy-driven scheduling features to optimize compute infrastructure resources and application performance
- Multicluster scheduling capability

ENEAGRID users accounting data (2)

CRESCO accounting data

- Data stored in *files* (LSF format)
- Two data types:
  - Job submissions
  - Logins
- *Raw* data store *every event* and/or change during job life

- ENEA developed ad-hoc preprocessing routines to produce a single summarized file per year with job info:
  - Status
  - Times (start – end)
  - Queue name
  - User
  - Submission frontend
  - … other LSF params…
- User login data (auth requests)

ENEAGRID users accounting data (3)

CRESCO accounting data

- ***Example: 2016***
  - **More than 850’000 job entries** in the aggregate file
  - User authentication data:
    - ENEAGRID site: Portici
    - Login server: afsdb.portici.enea.it (1 of 2 auth servers)
    - **#auth_req > 8’000’000 !!!**

Data correlation: consumption vs cores (1)

LSF and Zabbix data correlation

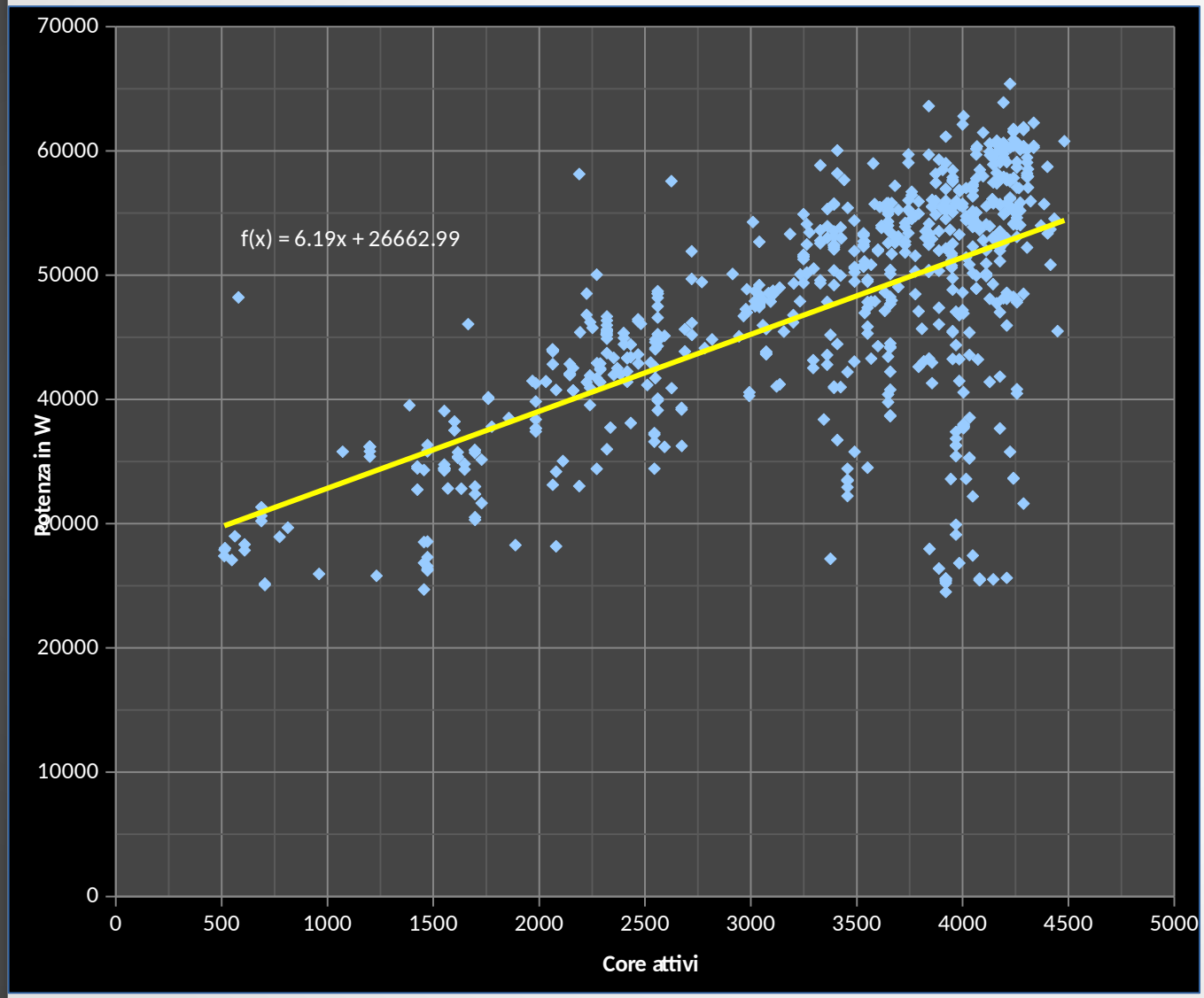
ZABBIX → **Power info**

1454889600	36040,00	37013,00	38310,00
1454893200	35460,00	36888,00	38160,00
1454896800	31610,00	34122,00	37600,00
1454900400	32640,00	34987,00	35840,00
1454904000	33720,00	35103,00	35880,00
1454907600	33400,00	35118,00	36080,00
1454911200	33120,00	35099,00	36050,00
1454914800	32680,00	34448,00	35980,00
1454918400	27830,00	32400,00	38590,00

LSF → **# Core info**

1	823895	cresco4	1454886067	1454891098	2671
2	823895	cresco4	1454886067	1454891098	2672
3	823895	cresco4	1454886067	1454891098	2672
4	823895	cresco4	1454886067	1454891098	2671
1	823896	cresco4	1454886072	1454896906	2671
2	823896	cresco4	1454886072	1454896906	2672
3	823896	cresco4	1454886072	1454896906	2672
4	823896	cresco4	1454886072	1454896906	2672

Data correlation: consumption vs cores (2)



Period: 6 months

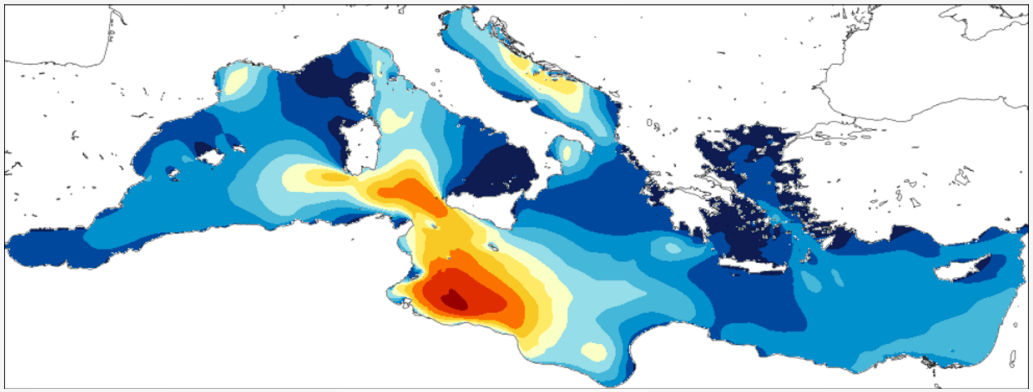
- Data form LSF and Zabbix have to be **aligned** to common timestamps
- #cores in time interval is the sum of the job active in the interval

- LSF data: 188’564
- Zabbix data: 10’686

Climate forecast (1)

Meteorology in ENEA

System to provide climate forecasts of the Mediterranean area (next 5 days)



Large data files are stored in NetCDF format:

- A consolidate standard for scientific data
- Self-describing, machine-independent data formats
- Allows to reduce data occupancy up to 7 times

Climate forecast (2)

Data storing and elaboration issues

Simulations are executed ***every day on CRESCO*** HPC cluster

- Input: ~100GB
- Output: ~500GB

Data grow up every day!

Four main issues for such a Big Data scenario:

- ***Data elaboration***: ad-hoc file systems to efficiently handle big data processing
- ***Data storing***: historical data are “gold value”
- ***Data visualization***: possibility to access end efficiently visualize forecast data
- ***Data correlation***: typical machine learning task to discover patterns and similarities among regions → Deep Learning tools and frameworks

Air pollution models (1)

The MINNI project
(www.minni.org)

MINNI is an Integrated Modeling System, that is a model description of a **complex system** consisting of several interdependent and interconnected components, each of which describes individual system aspects.

MINNI consists of two main modeling systems:

- Atmospheric Model System (AMS): describes physico-chemical processes in the atmosphere
- GAINS-Italy (Greenhouse Gas - Air Pollution Interactions and Synergies): allows evaluation of impacts and costs

Air pollution models (2)

Data storing and elaboration issues

Simulations are executed on ***CRESCO*** HPC cluster

Models produce and elaborate a large amount of data file storede in CRESCO data storage system:

- More than 180TB of data stored in two storage systems located in CRESCO Portici
- More than 6.6ML of files!

Data grow up every day!

Main issues for such a Big Data scenario:

- ***Data elaboration***: ad-hoc file systems to efficiently handle big data processing
- ***Data storing***: huge amount of data
- ***Data correlation***: typical machine learning task to discover patterns and similarities among regions → Deep Learning tools and frameworks

Web Crawler (1)

Web Crawler

Tool to browse www systematically and download web contents.

Data are stored locally and processed to build indexes, statistics and to structure them

Web snapshots are typically stored incrementally and can be analyzed to discover changes, new contents and the evolution of the web

Application contexts:

- Intelligence & security
- Blog analysis
- User behaviors
- Marketing

Web Crawler (2)

ENEA Web Crawling

Framework to support web crawling and data analysis.

Developed with open source products.

Customized and fully-integrated within ENEAGRID.

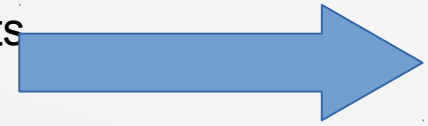
Exploits CRESCO storage, network, and computing resources

Experiments: periodical web snapshot

- Period: 1 month (August 2016)
- Crawling time: 1h
- Computing resources: 8 nodes, 16 agents (2 per node)

Results (per session):

- ~500GB of web contents
- ~11ML of web pages



Aggregate results

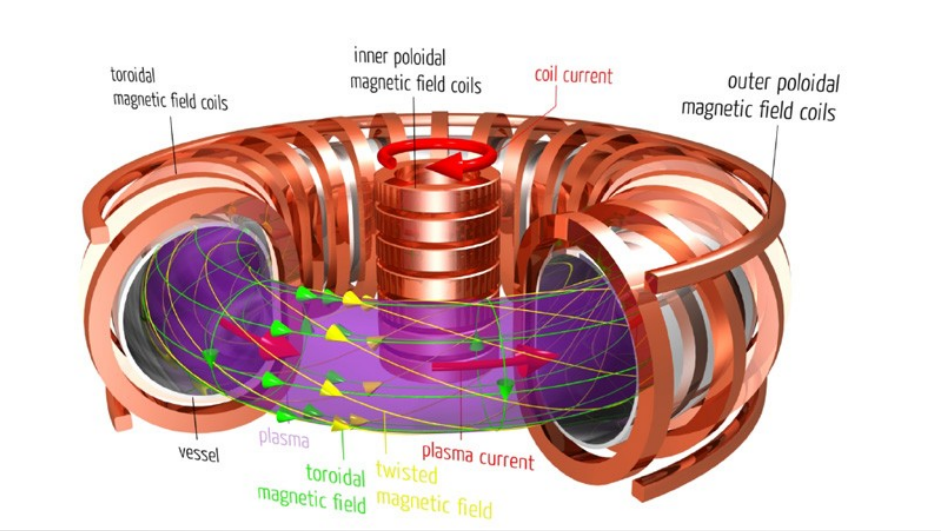
More than **15TB** of data and **~340ML** of web pages

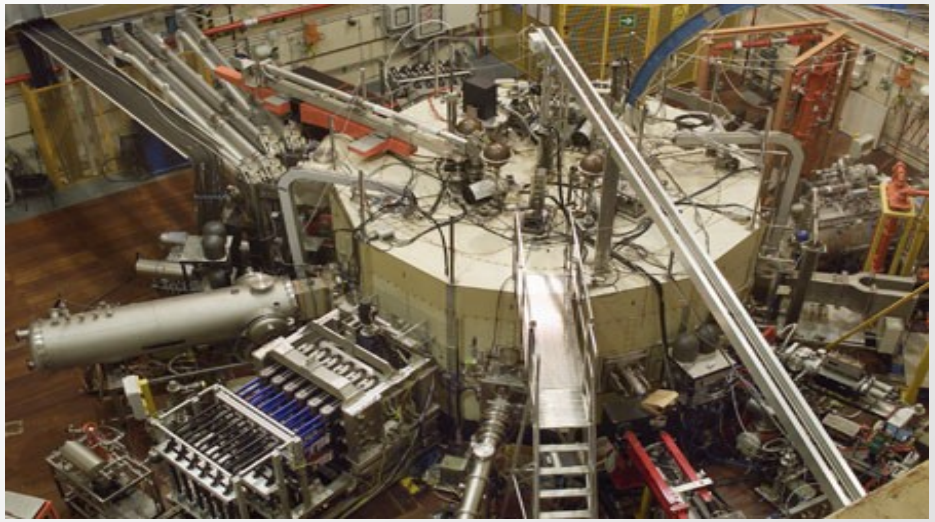
Nuclear fusion experiments (1)

Frascati Tokamak Upgrade (FTU) data

A Tokamak is a device that uses a powerful magnetic field to confine plasma in the shape of a torus.

A set of electromagnets polarize the plasma inducing an electric current inside.





Nuclear fusion experiments (2)

Experiment: plasma discharge data

- Data captured during plasma discharges (~2secs each)
- Several acquisition sampling (msecs → secs)
- Acquisition channels (diagnostic signals):
  - Simple raw data (f_t)
  - Complex multidimensional data

Big Data archive:

- Experimental data from 1989 to now
- About 40’000 plasma discharges
- ~2TB of raw data (not post-processed) stored in files
- Data files are in a standard format of the Joint European Torus (JET)
  - Each channel is stored in a file containing data and metadata

Big Data: what is more important?

- the “data”?
- the “big”?
- Both?
- …

Neither!

The crucial aspects is the “Information” and the “value”

The 4Vs + Value

Value: *Big data can generate huge competitive advantages*

Big Data Analytics

In order to extract “value” from big data, there should be adopted proper analytic tools.

Traditional machine learning tools should be adapted to face with Big Data issues

- Efficient in analyzing/mining data
- Do not scale

- Efficient in managing big data
- Not so easy to analyze or mine the data

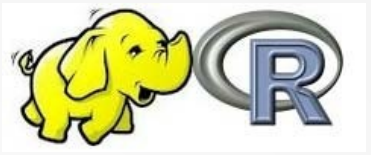
How to integrate them?

Big data projects

- R over a cluster computing framework
  - Rhadoop:

    Open source extension of R on Hadoop
  - Revolution R:

    R distribution from Revolution Analytics
  - ...
- Apache Mahout
  - Open-source package on Hadoop for data mining and machine learning
- Apache MLlib
  - Spark's scalable machine learning library consisting of common learning algorithms and utilities



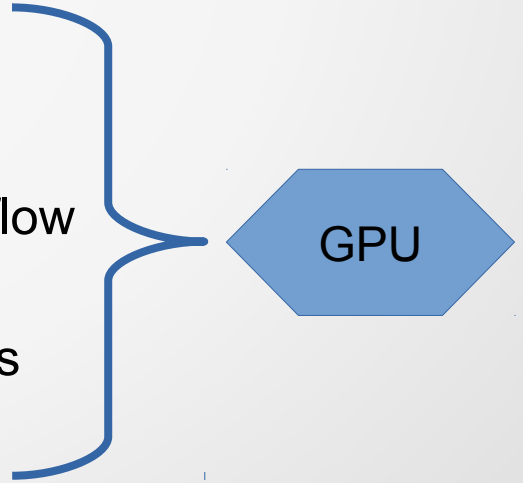
Deep Learning Frameworks

Artificial Intelligence field of self-learning exploiting machine learning algorithms in multiple hierarchical layers (nonlinear process).

Artificial neural networks are the main technique and framework for deep learning

Tools:

- ***Caffe***: Useful for feature extraction. Berkeley Univ.
- ***Torch***: wide support ML
- ***Tensor Flow***: library for numerical computation using data flow graphs. Developed by Google Brain team
- ***Theano***: Python lib for multi-dimensional array computations
- ***DL4J***: deep learning libraries for Java



Thanks for the attention!

Ing. Giovanni Ponti, PhD
ENEA – DTE-ICT-HPC

giovanni.ponti@enea.it