Mem. S.A.It. Suppl. Vol. 19, 324

Memorie della
Supplementi

DAME: A Distributed Web Based Framework for Knowledge Discovery in Databases

M. Brescia[1], G. Longo[2], M. Castellani[3], S. Cavuoti[2], R. D'Abrusco[4], and O. Laurino[5]

[1] Istituto Nazionale di Astrofisica – Osservatorio Astronomico di Napoli, Italy
[2] Dipartimento di Fisica, Università degli Studi Federico II, Napoli, Italy
[3] Istituto Nazionale di Astrofisica – Osservatorio Astronomico di Roma, Italy
[4] Harvard Smithsonian Center for Astrophysics, Cambridge MA, USA
[5] Istituto Nazionale di Astrofisica – Osservatorio Astronomico di Trieste, Italy

Abstract. Massive data sets explored in many e-science communities, as in the Astrophysics case, are gathered by a very large number of techniques and stored in very diversified and often-incompatible data repositories. Moreover, we need to integrate services across distributed, heterogeneous, dynamic virtual organizations formed from the different resources within a single enterprise and/or from external resource sharing and service provider relationships. The DAME project aims at creating a distributed e-infrastructure to guarantee integrated and asynchronous access to data collected by very different experiments and scientific communities in order to correlate them and improve their scientific usability. The project consists of a data mining framework with powerful software instruments capable to work on massive data sets, organized by following Virtual Observatory standards, in a distributed computing environment. The integration process can be technically challenging because of the need to achieve a specific quality of service when running on top of different native platforms. In these terms, the result of the DAME project effort is a service-oriented architecture, by using appropriate standards and incorporating Cloud/Grid paradigms and Web services, that will have as main target the integration of interdisciplinary distributed systems within and across organizational domains.

Key words. Astroinformatics – data mining – distributed computing – virtual observatory – machine learning

1. Introduction

Modern scientific data mainly consist of huge datasets gathered by a very large number of techniques and stored in very diversified and often incompatible data repositories. In the e-science environment, it is considered as a critical and urgent requirement to integrate services across distributed, heterogeneous, dynamic "virtual organizations" formed by different resources within a single enterprise. The Astronomy and Astrophysics environment has become an immensely data rich field due to the evolution of detectors, telescopes and space instruments. Different astrophysics areas share the same basic requirement: to be able to deal

Send offprint requests to: M. Brescia
e-mail: `brescia@oacn.inaf.it`

with massive and distributed datasets whereas possible integrated with services. This new understanding includes knowing how to access, retrieve, analyze, mine and integrate data from disparate sources. But on the other hand, it is obvious that a scientist cannot and does not necessarily want to become an expert in the fields of IT (Information Technology). The idea behind projects like DAME, described in this paper, is to provide a user friendly and standardized scientific gateway to easy the access, exploration, processing and understanding of massive data sets. In the field of astronomy, DAME represents a typical product of Knowledge Discovery in Databases (KDD) that is being recognized as the fourth leg of scientific research after theory, experiments and simulations. It arises from the pressing need to acquire the multi- disciplinary expertise which is needed to deal with the ongoing burst of data complexity and to perform data mining and exploration on Massive Data Sets (MDS). The data interoperability standardization approach in Astrophysics have been resulted in the Virtual Observatory (VO). It consists into the federation under common standards of all astronomical archives available worldwide. DAME has as one of main goals to extend and integrate the VO experiment, by making feasible the full interoperability between data and applications for analysis, mining and exploration. The main drive behind such effort being that once the infrastructure will be completed, it will allow a new type of multi-wavelength, multi-epoch science which can only be barely imagined.

2. Massive Data Mining in Astrophysics

An important part of the computing challenges in astronomy are related to the handling, processing and modeling of large quantities of data. All astrophysical carriers have their peculiarities and weaknesses from the scientific point of view: they sample different energy ranges, endure different kinds and levels of interference during their cosmic journey, sample different physical phenomena (e.g. thermal, non thermal and stimulated emission mechanisms), and require very different technologies for their detection. So far, the international community needs modern infrastructures for the exploitation of the ever increasing amount of data (of the order of PetaByte/year) produced by the new generation of telescopes and space borne instruments, as well as by numerical simulations of exploding complexity. So far, the basic requirements (extensible also to other application fields) can be summarized in two items: (a) the need of a federation of experimental data, by collecting them through several worldwide archives and by defining a series of standards for their formats and access protocols; (b) the implementation of innovative computing tools for Data Mining (DM) and knowledge extraction, user-friendly, scalable and as much as possible asynchronous; These topics require powerful, computationally distributed and adaptive tools able to explore, extract and correlate knowledge from multivariate massive datasets in a multi-dimensional parameter space. Every time a new technology enlarges the parameter space or allows a better sampling of it, new discoveries are bound to take place. So far, the scientific exploitation of a multi-band (D bands), multi-epoch (K epochs) universe implies to search for patterns and trends among N points in a DxK dimensional parameter space, where $N > 10^9$, $D >> 100$, $K > 10$. The problem also requires a multi-disciplinary approach, covering aspects belonging to Astronomy, Physics, Biology, Information Technology, Artificial Intelligence, Engineering and Statistics. As for data, the concept of distributed archives is already familiar to the average astrophysicist. The leap forward in this case is to be able to organize the data repositories to allow efficient, transparent and uniform access: these are the basic goals of the VO (Smareglia et al. 2006). In more than a sense, the VO is an extension of the classical Computational Grid. DAME extends this fundamental target by integrating it in an infrastructure, joining CLOUD service-oriented software and GRID resource-oriented hardware paradigm, including the implementation of advanced tools for MDS exploration (Brescia et al. 2009a; Brescia et al 2009b) (Fig. 2). In particular, concerning the GRID side, the Suite exploits the S.Co.P.E. GRID infrastructure. The S.Co.P.E. project (Merola 2008), aimed at the construction and activation of a Data Center which is now perfectly integrated in the national and international GRID initiatives, hosts 300 eight-core blade servers and 220 Terabyte of storage. The acronym stands for Cooperative System for Multidisciplinary Scientific Computations, that is a collaborative system for scientific applications in many areas of research. Moreover to overcome the limitation due to synchronous run of services, one of the main DAME design strategies is to permit asynchronous access to the infrastructure tools, allowing running of activity jobs and processes outside the scope of any particular web-service operation and without depending on the user connection status. The user, via client web applications, can asynchronously find out the state of the activity, has the possibility to keep track of his jobs by recovering related information (partial/complete results) without having the need to maintain open the communication socket. Furthermore, the DAME design takes into account the fact that the average scientists cannot and/or does not want to become an expert also in Computer Science. In most cases the r.m.s. scientist already possesses his own algorithms for data processing and analysis and has implemented private routines/pipelines to solve specific problems. These tools, however, often are not scalable to distributed computing environments. DAME aims at providing a user friendly web based tool to encapsulate own algorithm/procedure into the package, automatically formatted to follow internal programming standards. In our project data mining is intended as techniques of exploration on data, based on the combination between parameter space filtering, machine learning, soft computing techniques associated to a functional domain. The functional domain term arises from the conceptual taxonomy of research modes applicable on data. Dimensional reduction, classification, regression, prediction, clustering, image segmentation are example of functionalities belonging to the data mining conceptual domain, in which the various methods (models and algorithms) can be applied to explore data under a particular aspect, connected to the associated functionality scope. Such DM models are mainly derived from Machine Learning and Artificial Intelligence taxonomy: Multi Layer Perceptron (MLP) with classical Back Propagation or Genetic Algorithms learning paradigms; Support Vector Machine (SVM); Self-Organizing Maps, Principal Probabilistic Surfaces (PPS). All these analytical methods based partially on statistical random choices (crossover/mutation) and on knowledge experience acquired (supervised and/or unsupervised adaptive learning) could realistically achieve the discovery of hidden laws behind focused phenomena, often based on nature laws, therefore the simplest.

3. The Available Resources and Services

As previously remarked, DAME is organized as a CLOUD of web applications and services. The following are the ones already available and accessible from the project website (http://dame.dsf.unina.it):

DM WEB Application Prototype: a simplified model of main Suite web application, providing tools to configure and launch scientific experiments; SDSS (Sloan Digital Sky Survey) local site: local mirror website hosting a complete SDSS data archive and management system; WFXT (Wide Field X-ray Telescope) transient calculator: a web application to estimate the number of variable sources detected within the 3 main planned extragalactic surveys, with a given significant threshold; DM Web Application SUITE: the main service of the project, providing via browser a complete DM framework including dataset and experiment configuration, execution and graphical/text tools for outputs. The last release, under technical commissioning, is now available at: http://dame.dsf.unina.it/beta_info.html. To access it, an account is required (you can ask it to authors). The main page, reported in Fig. 3, provides utilities to organize user own workspaces, experiments and data, build input datasets, select experiment functionalities (at the moment only classifi-

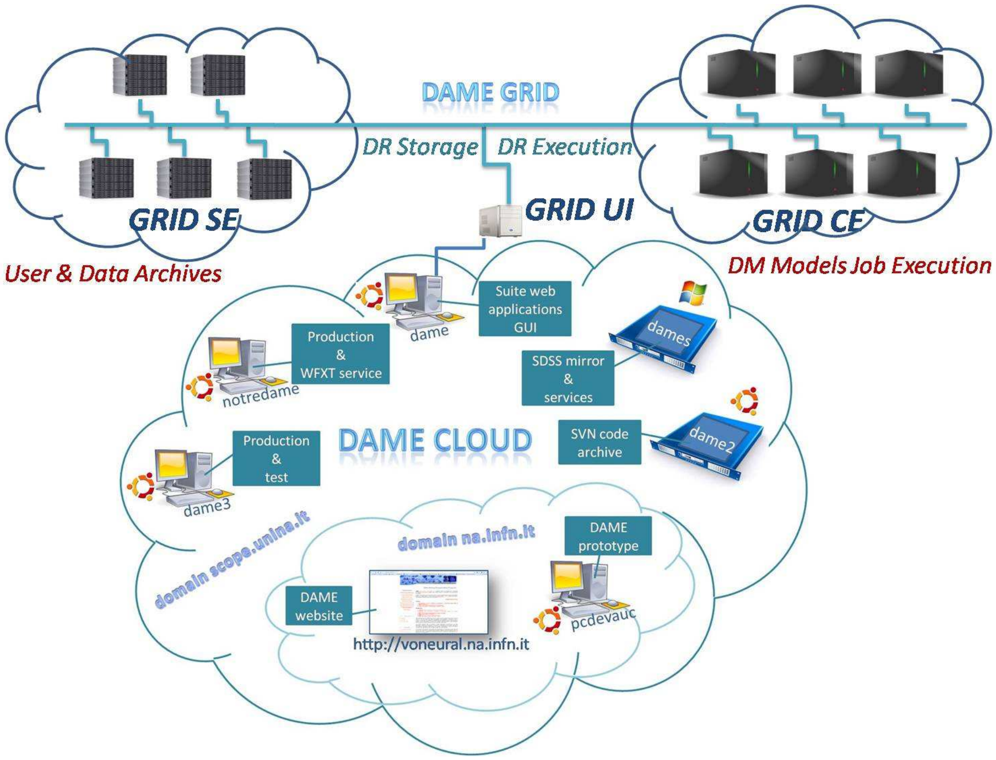

Fig. 1. The DAME hybrid infrastructure

cation/regression), select and configure DM model parameters (at the moment between MLP with Back Propagation or with Genetic Algorithms, SVM), execute jobs, download text and graphical results. In practice, the alpha release is almost complete in terms of available resources and services. Its basic role is of course to be subject of intensive tests, in order to validate the Suite in terms of usability, performance and reliability.

In the next Autumn it is also foreseen the releasing of other tools (currently under commissioning phase):

– **The public Beta release of DAME SUITE**, enhanced with more DM resources (clustering and image segmentation as new functionalities and some new DM models: Multi Layer Clustering, Self-Organizing Maps, MLP with Quasi Newton statistical method, based on L-BCFGS algorithm, and Probabilistic Principal Surfaces);
– **VOGCLUSTERS**: a VO-compliant web application for data and text mining on globular clusters; DM Web Application SUITE: the main service of the project, providing via browser a complete DM framework including dataset and experiment configuration, execution and graphical/text tools for outputs).

4. Scientific Impact and First Results

The models and algorithms provided with the Suite have been already tested on astrophysical use cases, with successful results, as reported in D'Abrusco et al. (2007) and D'Abrusco et al. (2009). The following are some examples.

i) **Photometric redshifts for the SDSS galaxies**

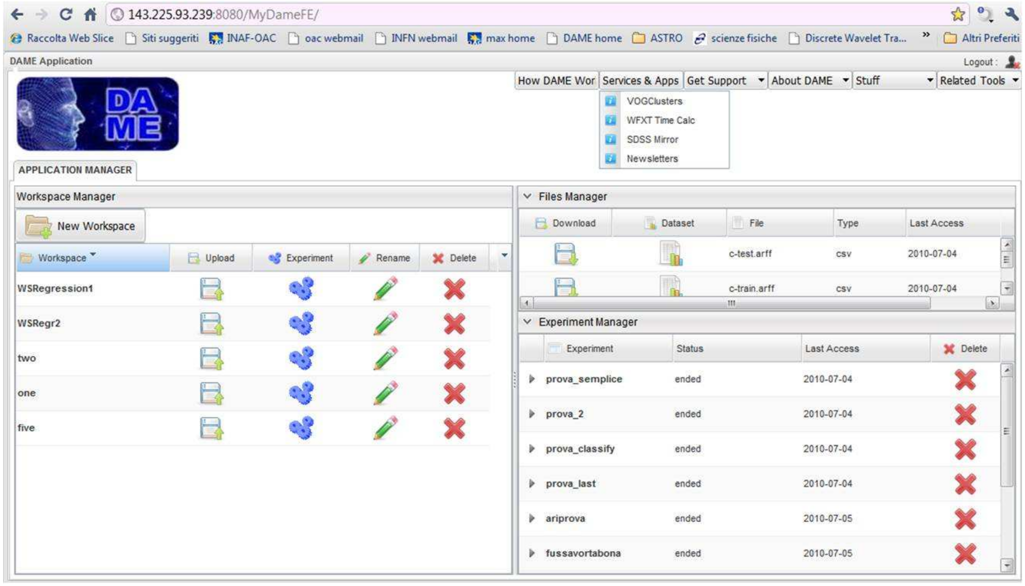

Fig. 2. The DAME alpha release

It makes use of a nested chain of MLP and allowed to derive the photometric redshifts for ca. 30 million SDSS galaxies with an accuracy of 0.02 in redshift. This result was also crucial for a further analysis of low multiplicity groups of galaxies (Shakhbazian) in the SDSS sample.

ii) **Search for candidate quasars in the SDSS**
The work was performed using the PPS module applied to the SDSS and SDSS+UKIDS data. It consisted in the search for candidate quasars in absence of a priori constrains and in a high dimensional photometric parameter space.

iii) **AGN Classification in the SDSS**
Using the DAME S.Co.P.E. GRID infrastructure to execute 110 jobs on 110 processors, the SVM model was employed to produce a classification of different types of AGN using the photometric data from the SDSS and the base of knowledge provided by the SDSS spectroscopic subsamples (a paper on this topic is in preparation).

iv) **DAME as an efficient computing service for DM experiments**
Concerning the computing results achieved using the hybrid architecture, it is possible to execute simultaneous experiments that gathered all together, bring the best results. Even if the single job is not parallelized, we obtain a running time improvement by reaching the limit value of the Amdahls law. For example, in the case of AGN Classification experiment (cited above), each of the 110 jobs runs for about a week on a single processor. By exploiting the GRID, the experiment running time can be reduced to about one week instead of more than 2 years (110 weeks).

v) **DAME as data mining use case at VO-DAYS**
From December 2009 to April 2010, the DAME web application prototype and the photometric redshifts case has been officially included in the four scientific use cases used in the interactive demonstration sessions, made at all INAF italian institutes, with unforeseen positive results (http://wwwas.oats.inaf.it/voday):

- 12 Sessions + 1 Videoconf with TNG, touching all cities with INAF structures;
- 6 tutors for each session (11 people involved);
- Registered: 272 (more than 1/4 of INAF research staff);
- Attendant: 244;
- Evaluation Form: 176 (forms are available at VO-day pages);
- About 70% already known VO as name (mainly they know VO tools but without using them);
- Several People request more specific tutorials on VO tools and how to publish own data on VO;
- About DAME use case, the 90% of users gave a positive feedback, evaluating as interesting and powerful the platform and considering the service "normal" in terms of hardness;

5. Conclusions

We have designed the DAME infrastructure to empower those who are not machine learning experts to apply these techniques and who have not proper resources to make own scientific experiments. One of the main goals of our approach is to contribute to the full interoperability between data (already obtained within IVOA) and applications (by following the forth paradigm of Science). Moreover, if extended to other scientific or applied research disciplines, the opportunity to gain new insights on the knowledge will depend mainly on the capacity to recognize patterns or trends in the parameter space, which are not limited to the 3-D human visualization, from very large datasets. In this sense DAME approach can be easily and widely applied to other scientific, social, industrial and technological scenarios. By its nature, DAME is an open and incremental project, offering intrinsic educational and expertise formation opportunities in the Astrophysics and IT research fields. Our project has recently passed the R&D phase, de facto entering in the commissioning step. But first scientific test results already confirm the goodness of the theoretical approach and technological strategy.

Acknowledgements. The DAME project run jointly by the Department of Physics of the University Federico II, INAF (National Institute of Astrophysics) Astronomical Observatory of Napoli, and the California Institute of Technology, is financed through grants from the Italian Ministry of Foreign Affairs, the European projects VO-TECH and VO-AIDA and by the USA - National Science Foundation. DAME makes use of distributed computing environments (e.g. the S.Co.P.E. - GRISU infrastructure) and matches the international IVOA standards. The authors thank all members of project working group and all DAME contributors.

References

Brescia, M., et al., 2009, Mem S.A.It. Suppl., Vol 13, 56

Brescia, M., et al., 2009, Final Workshop of GRID Projects P̈ON Ricerca 2000-2006, Avviso 1575", Catania, Italy

D'Abrusco, R., et al. 2007, ApJ, 663, 752

D'Abrusco, R., Longo, G., Walton, N.A., 2009, MNRAS, 396, 223

Merola, L., 2008, The SCOPE Project. Proceedings of the FINAL WORKSHOP OF GRID PROJECTS PON RICERCA 2000-2006, AVVISO 1575. Catania, Italy

Smareglia, R., et al., 2006, Mem. S.A.It. Suppl., Vol 9, 423