DAME: a distributed data mining infrastructure for e-science discoveries

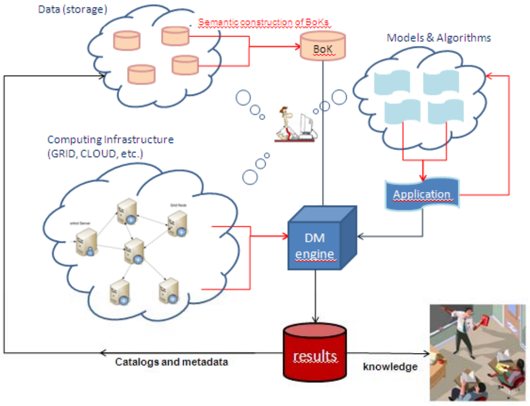
http://voneural.na.infn.it/

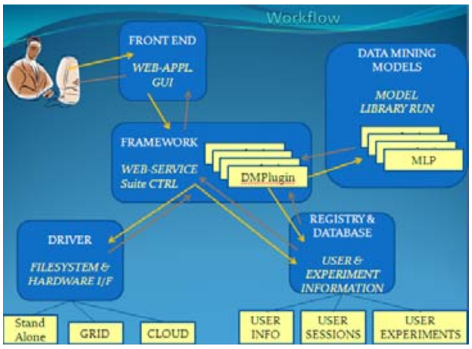
M. Brescia[1], R. D'Abrusco[2], C. Donalek[3], S.G. Djorgovski[3], O. Laurino[2], G. Longo[2], A. Mahabal[3], S. Cavuoti[2], G. d'Angelo[2], M. Garofalo[2], A. Nocella[2]

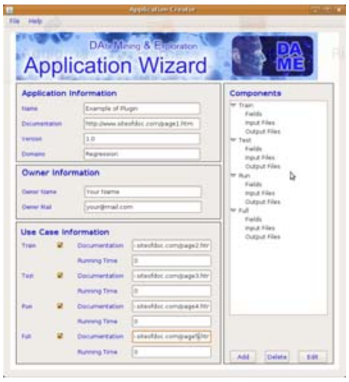
1 - INAF – Osservatorio Astronomico di Capodimonte, Napoli, Italy; 2- Department of Physics, University Federico II, Napoli, Italy;
3 - Department of Astronomy, California Institute of Technology, Pasadena, USA

In many e-science communities and environments Massive Data Sets (MDS) are gathered by means of heterogeneous techniques and stored in very diversified and often-incompatible data repositories. Moreover, it is needed to integrate services across distributed, heterogeneous, dynamic "virtual organizations" formed by resources available within a single enterprise and/or from external resource sharing and service provider relationships. The **DAME** project aims at creating a general purpose Data Mining and Exploration e-infrastructure capable to ensure integrated and asynchronous processing of data stored in such MDS.

The project consists of a data mining framework with powerful software tools capable to work on massive data sets, compliant with 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.







Application Information

Name: Example of Plugin
Documentation: http://www.sitesofdoc.com/page1.htm
Version: 1.0
Domains: Regression

Owner Information

Owner Name: Your Name
Owner Mail: your@mail.com

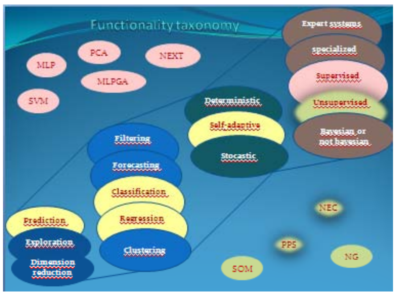
Use Case Information

Train ☑ Documentation: ...sitesofdoc.com/page2.htm; Running Time: 0
Test ☑ Documentation: ...sitesofdoc.com/page3.htm; Running Time: 0
Run ☑ Documentation: ...sitesofdoc.com/page4.htm; Running Time: 0
Full ☑ Documentation: ...sitesofdoc.com/page5.htm; Running Time: 0

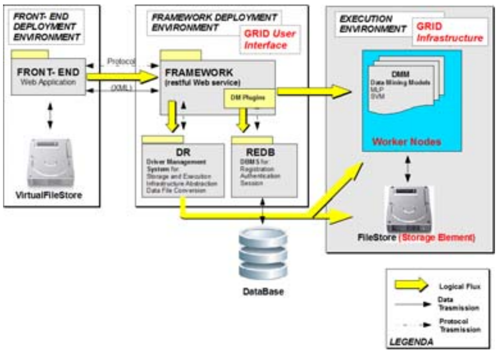
The project is based on five main components: Front End (FE), Framework (FW), Registry and Data Base (REDB), Driver (DR) and Data Mining Models (DMM), plus an application plugin wizard for source code generation

It is a web oriented and VO aware suite. In order to perform data mining and exploration experiments, we have considered a top-down strategy, starting from the taxonomy of data mining and research functionalities which are associated to specific algorithms and processing methods. In the first release, the suite offers tools for the following functionalities:

- ❖ **Classification**
- ❖ **Regression**
- ❖ **Clustering**

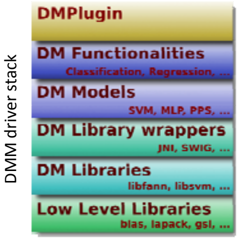


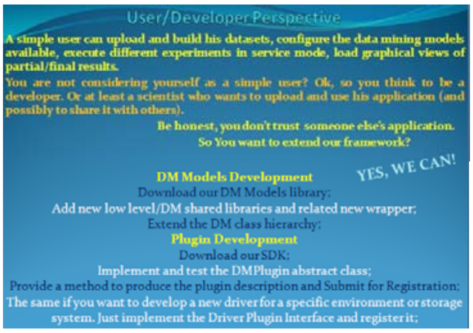
The above picture shows both already implemented models and next foreseen release evolutions



Main features of the complete package:

- ✓Object Oriented Programming & UML
- ✓Internal standards and protocols (XML)
- ✓Java language (generic for DMM)
- ✓User/Session Registry DB (MySQL)
- ✓Web-based User I/O (GWT-Ext)
- ✓Web Application and Web Service Technology
- ✓Plugin Modularity (easy to be integrated/modified)
- ✓Hardware independent through platform driver
- ✓Data conversion and manipulation support




User/Developer Perspective

A simple user can upload and build his datasets, configure the data mining models available, execute different experiments in service mode, load graphical views of partial/final results.

You are not considering yourself as a simple user? Ok, so you think to be a developer. Or at least a scientist who wants to upload and use his application (and possibly to share it with others).

Be honest, you don't trust someone else's application.
So You want to extend our framework?

YES, WE CAN!

DM Models Development
Download our DM Models library;
Add new low level/DM shared libraries and related new wrapper;
Extend the DM class hierarchy;

Plugin Development
Download our SDK;
Implement and test the DMPlugin abstract class;
Provide a method to produce the plugin description and Submit for Registration;
The same if you want to develop a new driver for a specific environment or storage system. Just implement the Driver Plugin Interface and register it;

DAME - DAta Mining and Exploration
DAME - DAta Mining and Exploration
DAME prototype