Tools and Services for Distributed Knowledge Discovery on Grids

Domenico Talia

(Joint work with **Mario Cannataro** and **Paolo Trunfio**)

DEIS

University of Calabria, ITALY

talia@deis.unical.it

HPC 2002, Cetraro, June 2002

OUTLINE

- Introduction
- Parallel and Distributed Data Mining on Grids
- Models, Prototypes, and Projects
- The KNOWLEDGE GRID
  - KNOWLEDGE GRID Architecture
  - KNOWLEDGE GRID Services
  - KNOWLEDGE GRID Tools
  - VEGA
- Conclusion

INTRODUCTION

DATA MINING and **KNOWLEDGE DISCOVERY**:

The process of discovering valid, novel, useful, and understandable patterns or models in

Massive Databases.

Several mining tasks:

- Classification,
- Clustering,
- Association,
- Episode discovery
- ...

Several application areas:

- life sciences
- physics
- geology
- e-commerce
- ...

INTRODUCTION

WHY DATA MINING ?

- Lots of data collected and warehoused.
- Data collected and stored at enormous speeds in local databases, from remote sources, or from the sky.
- Scientific simulations generating terabytes of data.

- **Huge data sets are hard to understand**.
- Traditional techniques are infeasible for raw data.

- Data mining helps
  - scientists in hypothesis formation in biology, medicine, physics, geology, engineering, ...
  - companies to provide better, customized services and support decision making.

PARALLEL AND DISTRIBUTED DATA MINING ON GRIDS

- When **large data sets** are coupled with **geographic distribution** of data, users and systems, it is necessary to combine different technologies for implementing **high-performance distributed knowledge discovery systems (PDKD)**.

- Grid middleware targets technical challenges in areas such as communication, scheduling, security, information, data access, and fault detection.

- Efforts are needed for the development of knowledge discovery tools and services on the computational grid.

Grid-aware PDKD systems

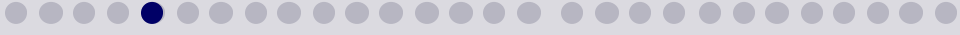
MODELS, PROJECTS and PROTOTYPES

- **KNOWLEDGE GRID**
  - **Discovery Net** — EPSRC's project at Imperial College (e-Science)
  - **DataCentric Grid** — Queen's University project/model for immovable data
  - **ADaM** — Algorithm Develop. and Mining to mine hydrology data
- **TeraGrid Project**
  - **Terra Wide Data Mining Testbed**
  - **Terabyte Challenge Testbed**
  - **Global Discovery Network**

Projects and Testbeds of the National Center for Data Mining (NCDM) at UIC.

KNOWLEDGE GRID

- **KNOWLEDGE GRID** - a PDKD architecture that integrates data mining techniques and computational grid resources.
- In the **KNOWLEDGE GRID** architecture **data mining tools** are compatible with lower-level **Grid mechanisms and services** and also with the **Data Grid services**.
- This approach benefits from "standard" grid services and offers an **open PDKD architecture** that can be configured **on top** of grid middleware.
- Grid infrastructure tools, such as Globus Tookit, provide basic services to be used in the development of the **KNOWLEDGE GRID**.



KNOWLEDGE GRID ENVIRONMENT

A **KNOWLEDGE GRID** computation uses:

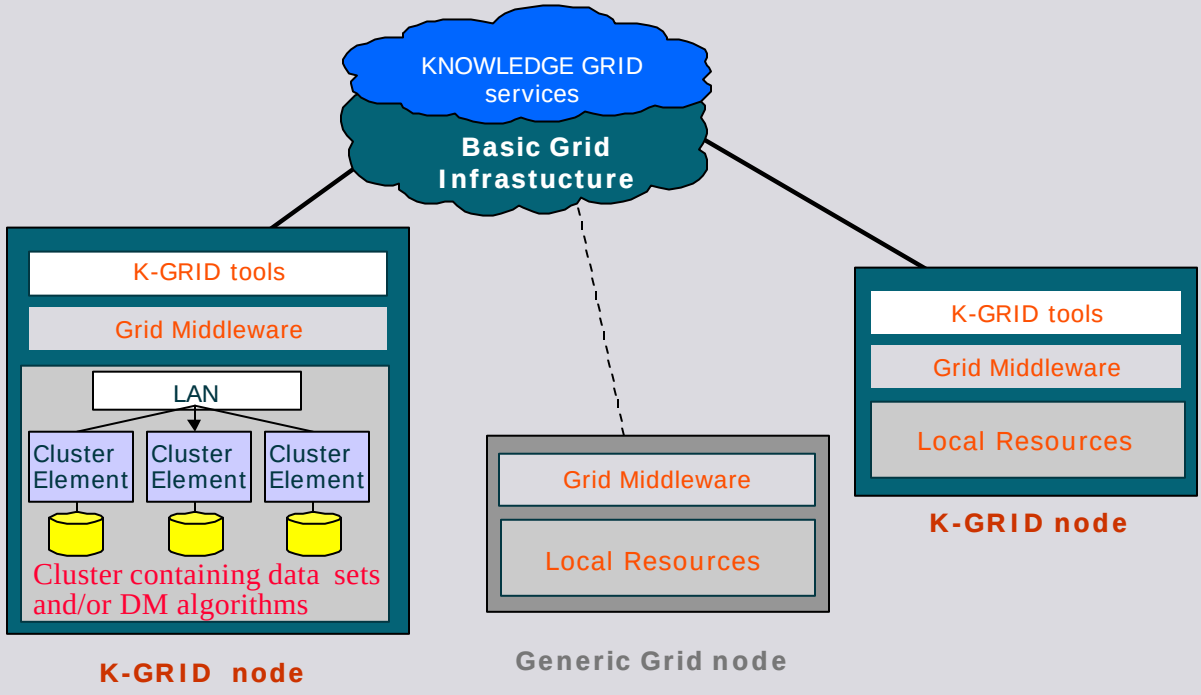
- **A set of KNOWLEDGE GRID-enabled computers - K-GRID nodes**

  declaring their availability to participate to some PDKD computation, that are connected by

- **A Grid infrastructure**

  offering basic grid-services (authentication, data location, service level negotiation) and implementing the KNOWLEDGE GRID services.

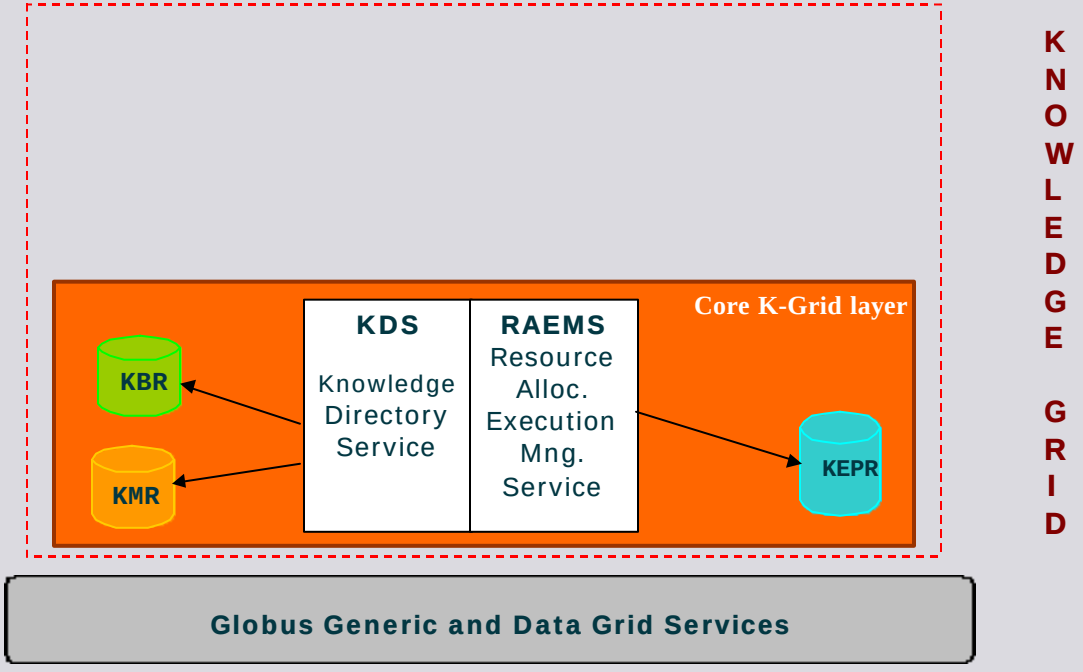
KNOWLEDGE GRID ENVIRONMENT



KNOWLEDGE GRID SERVICES

- The **KNOWLEDGE GRID** services are organized in two hierarchic layers :
  - *Core K-grid layer* and
  - *High-level K-grid layer*.
- The former refers to services directly implemented on the top of generic grid services.
- The latter is used to describe, develop and execute PDKD computations over the **KNOWLEDGE GRID**.

KNOWLEDGE GRID ARCHITECTURE



KNOWLEDGE GRID SERVICES

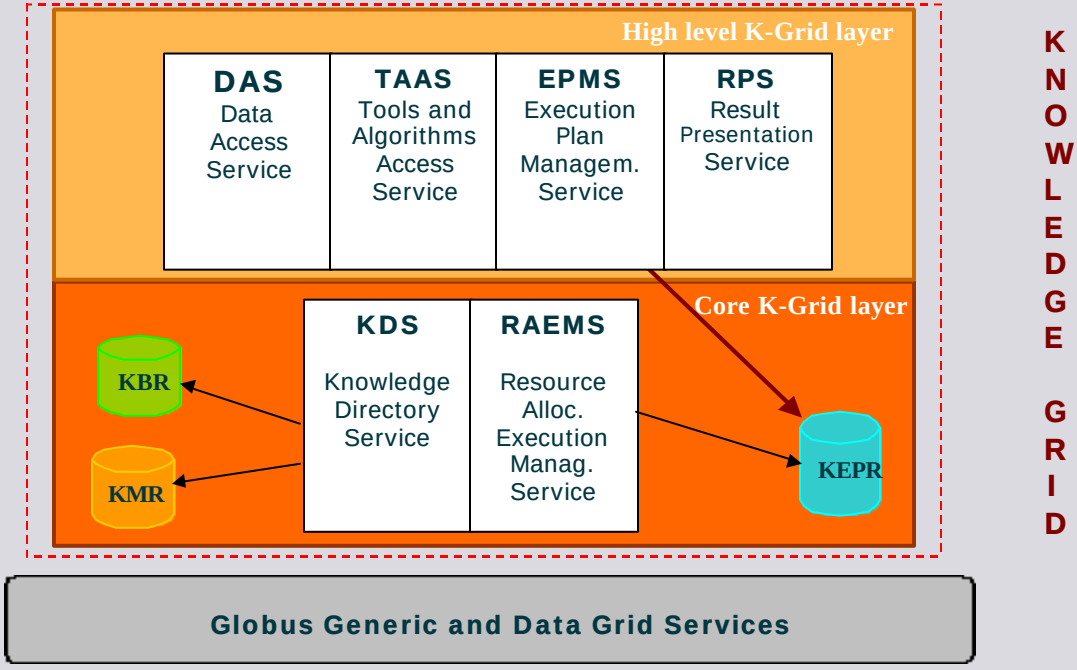
Core K-grid layer functions:

- Support of definition, composition and execution of a PDKD computation over the grid,
- Management of all metadata describing features of data sources, third party data mining tools, data management, and data visualization tools and algorithms.

Core K-grid layer services:

- **Knowledge directory service (KDS)**. Extends the basic Globus MDS and GIS services and maintains a description of all data and tools used in the **KNOWLEDGE GRID**.
- **Resource allocation and execution management service (RAEMS)**. RAEMS services are used to find a mapping between an execution plan and available resources.

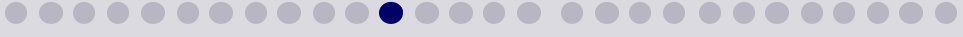
KNOWLEDGE GRID ARCHITECTURE



KNOWLEDGE GRID SERVICES

High-level K-grid layer services:

- **Data Access**
  - Search, selection (*Data search services*), extraction, transformation and delivery (*Data extraction services*) of data to be mined.
- **Tools and algorithms access**
  - Search, selection, and downloading of data mining tools and algorithms.
- **Execution Plan Management**
  - Generation of a set of different execution plans that satisfy user, data and algorithms requirements and constraints.
- **Results presentation**
  - Specifies how to generate, present and visualize the PDKD results (rules, associations, models, classification, etc.).



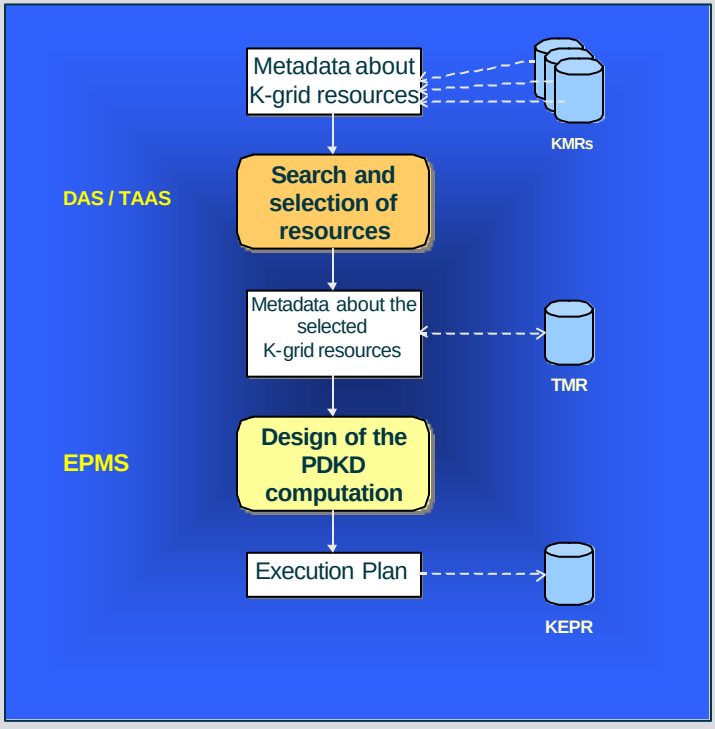
KNOWLEDGE GRID OBJECTS

- Metadata describing relevant **K-grid objects**, such as data sources and data mining tools, are implemented using both *LDAP* and *XML*.
- The (Knowledge Metadata Repository) **KMR** is implemented by LDAP entries and XML documents. The LDAP portion is used as a first point of access to more specific information represented by XML documents.
- LDAP object classes such as

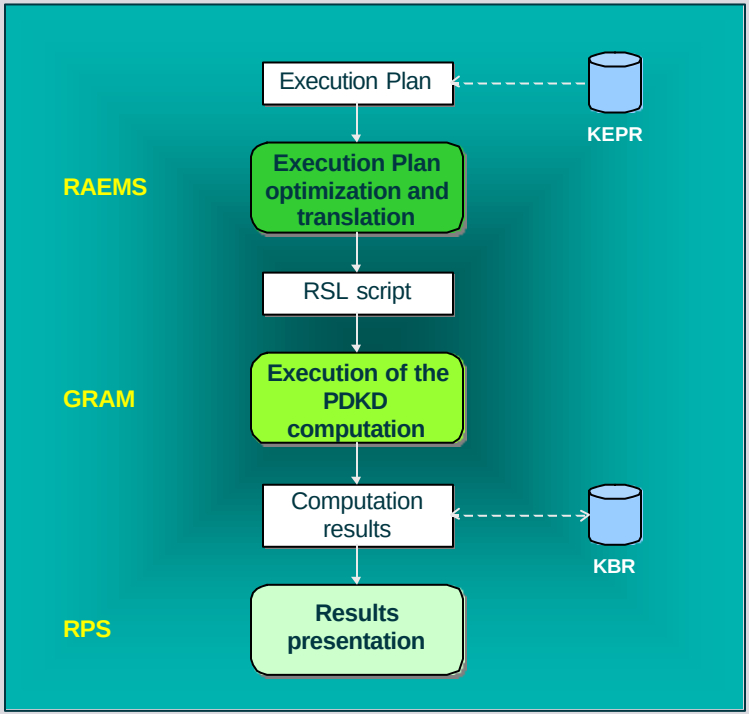
  K-gridDataSources and **K-gridSoftware**

  are used by a K-grid node to publish, respectively, the availability of data sources and software tools.

APPLICATION COMPOSITION STEPS



APPLICATION EXECUTION STEPS



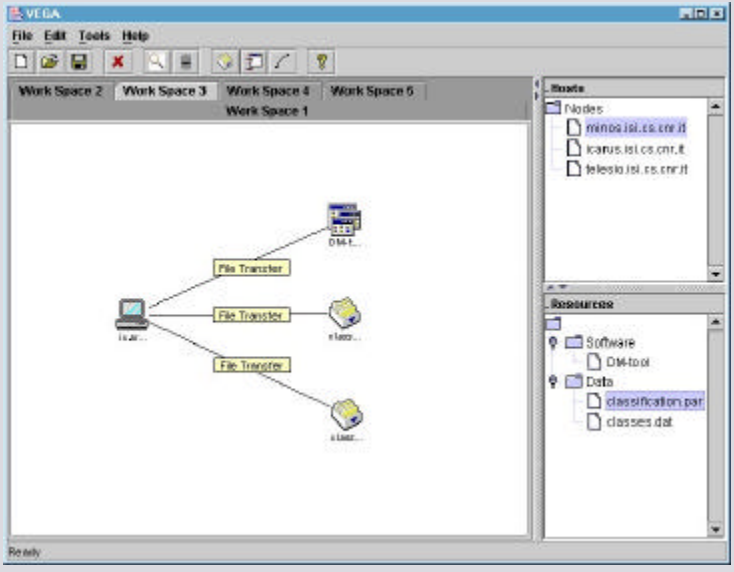
VEGA

- To allow a user to build a data mining application, we developed a toolset named **VEGA** (*a Visual Environment for Grid Applications*).

- **VEGA** offers support for :
  - task composition - definition of the entities involved in the computation and specification of the relations among them;
  - checking of the consistency of the planned task;
  - generation of the execution plan for a data mining task.
  - execution of the generated execution plan through the resource allocation manager of the underlying grid.

VEGA : OBJECTS and LINKS

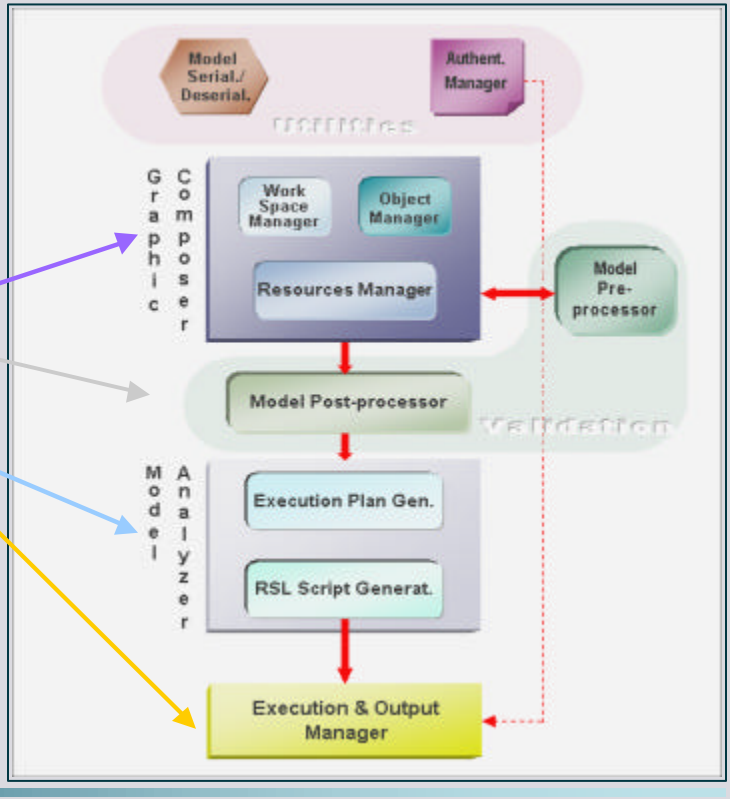
- **Objects**:
  - Hosts
  - Data
  - Software
- **Links**:
  - File Transfer
  - Execute
  - Input
  - Output



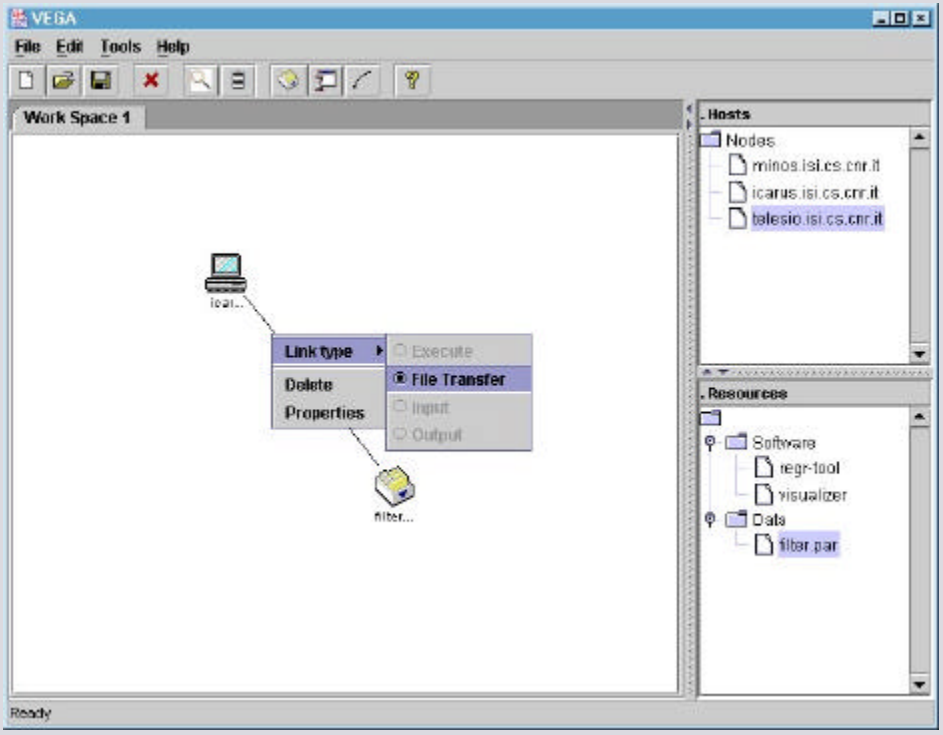
- *Objects* represent resources
- L*inks* represent relations among the resources

VEGA ARCHITECTURE

- Execution & Output Manager:
  - provides the graphical objects and functionalities for computation and manipulation
  - provides output to the user



VEGA

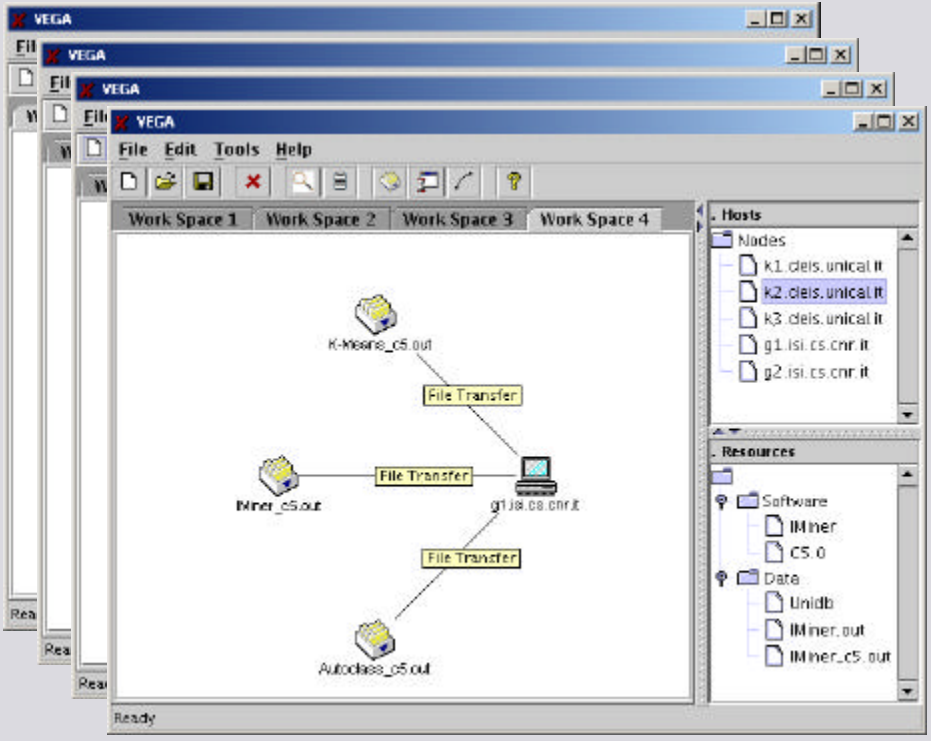


Hosts panel

Resources panel

WORKSPACES

A computation can be composed of several workspaces



XML METADATA in a KMR

```
...
<Software>
  <name>AutoClass</name>
  <description>Unsupervised Bayesian Classifier
   </description>
  <release>
    <number major="3" minor="3" patch="3"/>
    <date>01 May 00</date>
  </release>
  <author>Nasa Ames Research Center</author>
  <hostname>icarus.isi.cs.cnr.it</hostname>
  <executablePath>/share/software/autoclass-c/autoclass
   </executablePath>
  <manualPath>/share/software/autoclass-c/read-me.text
   </manualPath>
  ...
</Software>
```

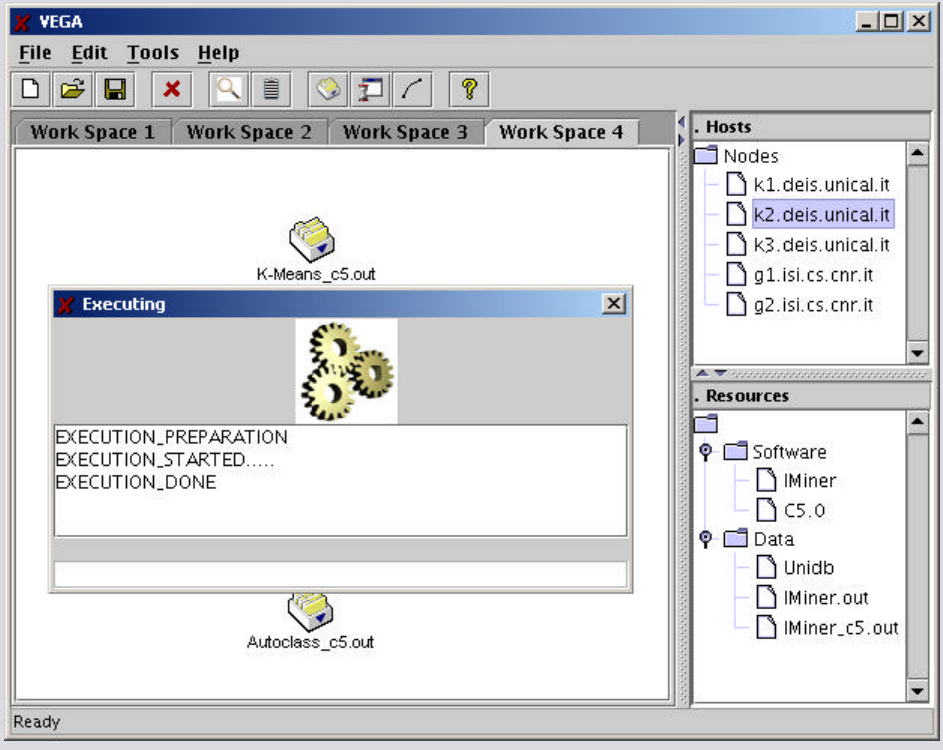
XML EXECUTION PLAN

```
<ExecutionPlan>
 ...
 <Task ep:label="ws1_dt2">
  <DataTransfer>
   <Source ep:href="g1../Unidb.xml" ep:title="Unidb on g1.isi.cs.cnr.it"/>
   <Destination ep:href="k2../Unidb.xml“ ep:title="Unidb on
      k2.deis.unical.it"/>
   ...
  </DataTransfer>
 </Task>
 ...
 <Task ep:label="ws2_c2">
  <Computation>
   <Program ep:href="k2../IMiner.xml" ep:title="IMiner on k2.deis.unical.it"/>
   <Input ep:href="k2../Unidb.xml" ep:title="Unidb on k2.deis.unical.it"/>
   ...
   <Output ep:href="k2../IMiner.out.xml" ep:title="IMiner.out on
      k2.deis.unical.it"/>
  </Computation>
 </Task>
 ...
 <TaskLink ep:from="ws1_dt2" ep:to="ws2_c2"/>
 ...
</ExecutionPlan>
```

A GENERATED RSL SCRIPT

```
+
...
(&(resourceManagerContact=g1.isi.cs.cnr.it)
  (subjobStartType=strict-barrier)
  (label=ws1_dt2)
  (executable=$(GLOBUS_LOCATION)/bin/globus-url-copy)
  (arguments=-vb –notpt gsiftp://g1.isi.cs.cnr.it/.../Unidb
             gsiftp://k2.deis.unical.it/.../Unidb
  )
)
...
(&(resourceManagerContact=k2.deis.unical.it)
  (subjobStartType=strict-barrier)
  (label=ws2_c2)
  (executable=.../IMiner)
  ...
  )
)
...
```

APPLICATION EXECUTION



FUTURE WORK

Some things to do on

VEGA :

- Support for more complex computation layouts,
- Execution plan optimization,
- Support for message passing (MPICH-G) applications.

KNOWLEDGE GRID :

- A peer-to-peer system for presence management and resource discovery on the Grid,
- A tool for optimized file transfer on the Grid based on GridFTP,
- Grid Portals: high-level **Problem Solving Environment (PSEs) for Knowledge Discovery on the Grid**.

CONCLUSIONS

- **Parallel and distributed data mining suites** and **computational grid technology** are two critical elements of future high-performance computing environments for
  - e-science (data-intensive experiments)
  - e-business (on-line services)
  - virtual organizations support (virtual teams, virtual enterprises)

- Knowledge grids will enable entirely new classes of **advanced applications** for dealing with the **data deluge**.

- Their integration is a challenge whose achievements could produce many benefits.

CONCLUSIONS

- Grids are **coupling** computation-oriented services with data-oriented services and high-level information management services.
- This trend enlarges the grid application scenario.
- We are much more able to store data than to extract knowledge from it.
- The **KNOWLEDGE GRID** allows for the

  unification of knowledge discovery and grid technologies

  helping us to climb some mountain of data.

THANKS