Distributed Data Mining Tasks and Patterns as Services

Large Grain Programming in Grids and Distributed Infrastructures

Domenico Talia

UNIVERSITY OF CALABRIA, Italy

talia@deis.unical.it

Goal

- Discuss a strategy based on the use of services for the design of open distributed knowledge discovery tasks and applications on Grids and distributed systems.

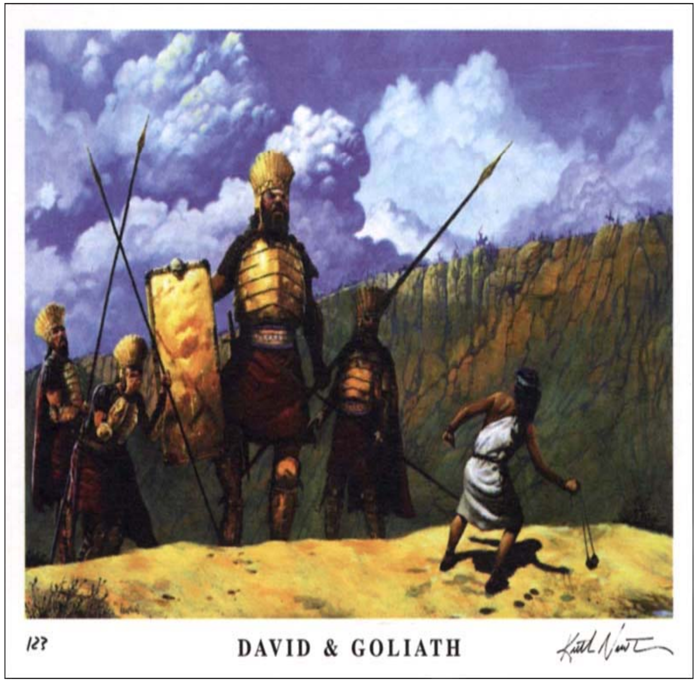
- Outline how Grid-based and service-oriented programming mechanisms can be developed as a collection of Grid/Web/Cloud services.

- Investigate how they can be used to develop distributed data analysis tasks and knowledge discovery applications exploiting the SOA model.

Complex Big Problems

- Bigger and more complex problems must be solved by distributed computing.

- DATA SOURCES are larger and larger and distributed.



- The main problem is not storing DATA, it is analyse, mine, and process DATA.

Data Availability or Data Deluge ?

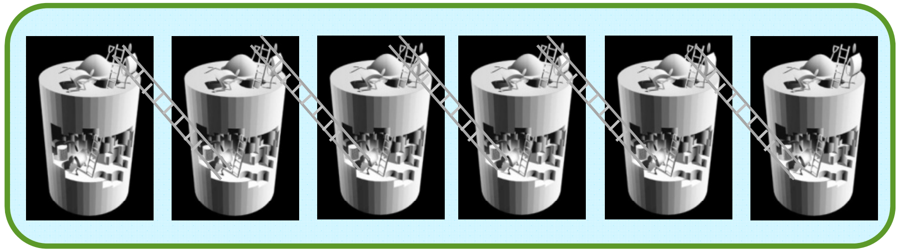
- Today the information stored in digital data archives is enormous and its size is still growing very rapidly.

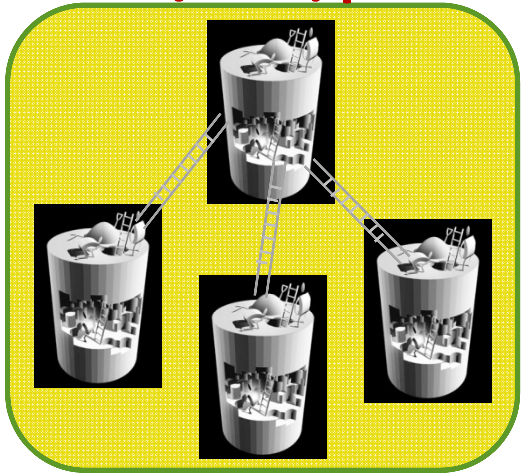
> WIRED
>
> The world has created or 161 exabytes (161 billion gigabytes) of digital information in 2006.
>
> (source: IDC)

- Whereas until some decades ago the main problem was the **shortage of information**, the challenge now seems to be
  - the **very large volume of information** to deal with and
  - the **associated complexity** to process it and to extract significant and useful parts or summaries.

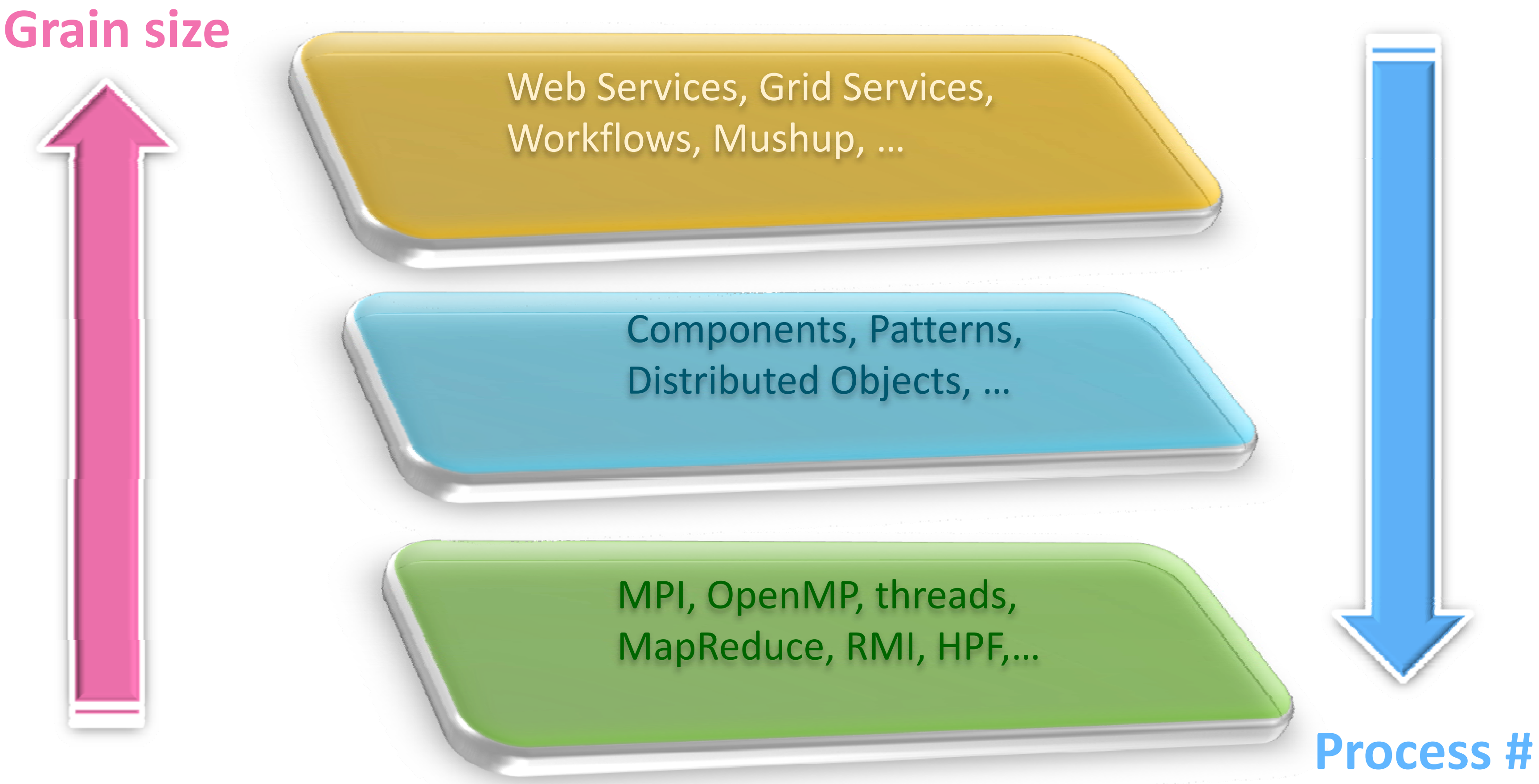
Distributed Data Analysis Patterns

- **Data parallelism? Task parallelism?**
- **Managing data dependencies**
- **Data management**: input, intermediate, output
- **Dynamic task graphs/workflows** (data dependencies)
- **Dynamic data access** involving large amounts of data
- **Parallel data mining** and/or **Distributed data mining**
- Programming **distributed mining operations/taks/patterns**





Programming Levels



Grain size

Web Services, Grid Services, Workflows, Mushup, …

Components, Patterns, Distributed Objects, …

MPI, OpenMP, threads, MapReduce, RMI, HPF,…

Process #

Distributed Data Mining on Grids

- The Grid extends the distributed and parallel computing paradigms allowing resource negotiation, dynamical allocation, heterogeneity, open protocols and services.
- As Grids and Clouds became well accepted computing infrastructures it is necessary to provide data mining services, algorithms, and applications.
- Those may help users to leverage Grid/Cloud/… capability in supporting high-performance distributed computing for solving their data mining problems in a distributed way.

Grid services for distributed data mining

- Exploiting the SOA model and the Web Services Resource Framework (WSRF) it is possible to define basic services for supporting distributed data mining tasks in Grids

- Those services can address all the aspects that must be considered in data mining and in knowledge discovery processes
  - data selection and transport services,
  - data analysis services,
  - knowledge models representation services, and
  - visualization services.

Grid services for distributed data mining

- It is possible to define services corresponding to

Single Steps

that compose a KDD process such as preprocessing, filtering, and visualization.

Single Data Mining Tasks

such as classification, clustering, and association rules discovery.

Distributed Data Mining Patterns

such as collective learning, parallel classification and meta-learning models.

Data Mining Applications or KDD processes

including all or some of the previous tasks expressed through a multi-step workflow.

Data mining Grid services

- This collection of data mining services can constitute an

Open Service Framework for Grid-based Data Mining

- Allowing developers to program distributed KDD processes as a composition of single and/or aggregated services available over a Grid.

- Those services should exploit other basic Grid services for data transfer and management for data tranfer, replica management, data integration and querying.

Data mining Grid services

- By exploiting the Grid services features it is possible to develop data mining services accessible every time and everywhere.

- This approach may result in
  - Service-based distributed data mining applications
  - Data mining services for virtual organizations.
  - Distributed data analysis services on demand.
  - A sort of knowledge discovery eco-system formed of a large numbers of decentralized data analysis services.

Data mining Grid services: Are they programming abstractions?

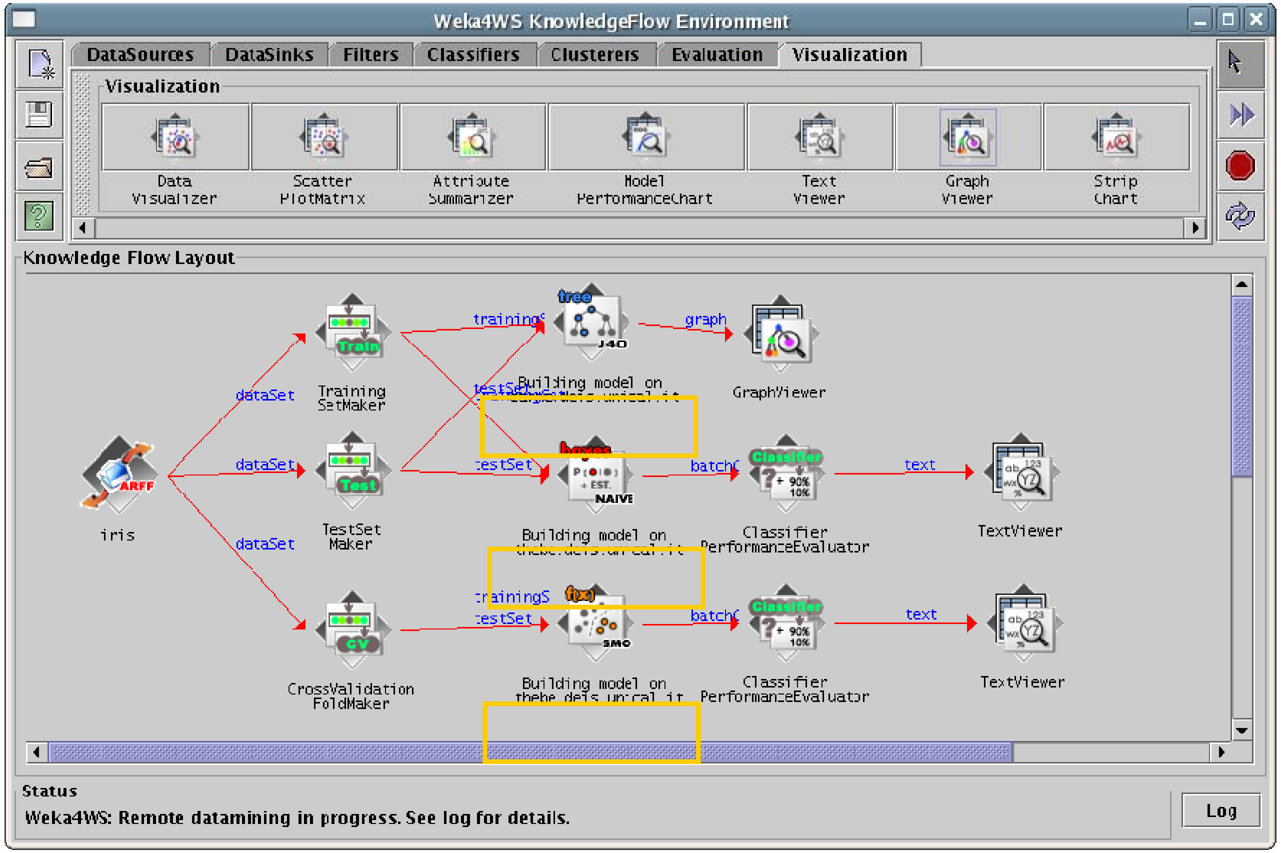
- Apparently **not**, in a traditional approach.

- **Yes**, if we consider the user and application requirements in handling data and in understanding what is useful in it.
  - Basic services as simple operations;
  - Service programming languages for composing them;
  - Complex services and their complex composition;
  - Towards distributed programming patterns for services.

Grid services for distributed data mining

- Service-based systems we developed
  - Weka4WS
  - Knowledge Grid
  - Mobile Data Mining Grid Services
  - Mining@home

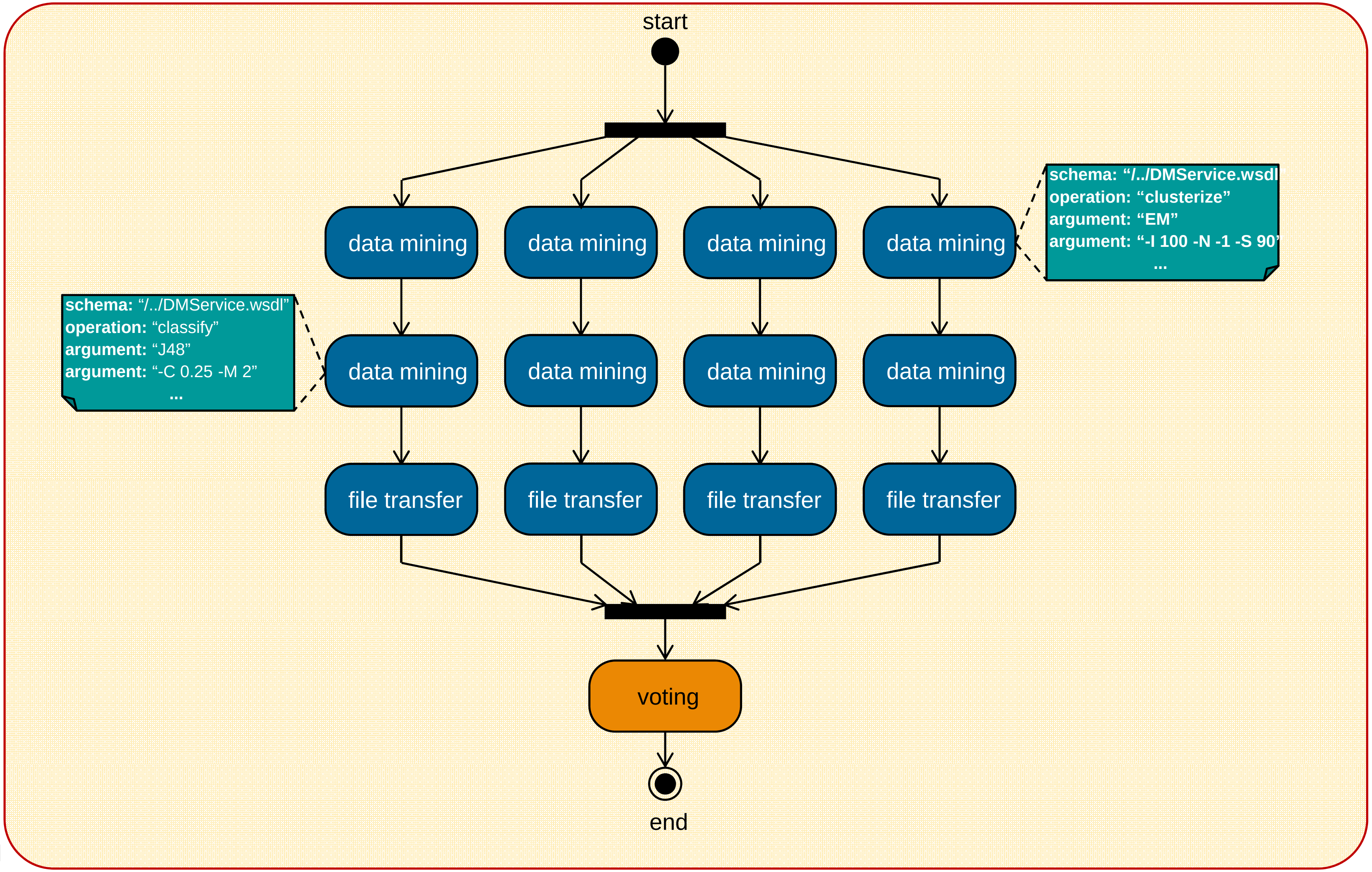
Weka4WS KnowledgeFlow



Weka4WS: Remote datamining in progress. See log for details.

Programming a data mining workflows and run them on several Grid nodes.

Knowledge Grid: application design



start

data mining data mining data mining data mining

data mining data mining data mining data mining

file transfer file transfer file transfer file transfer

voting

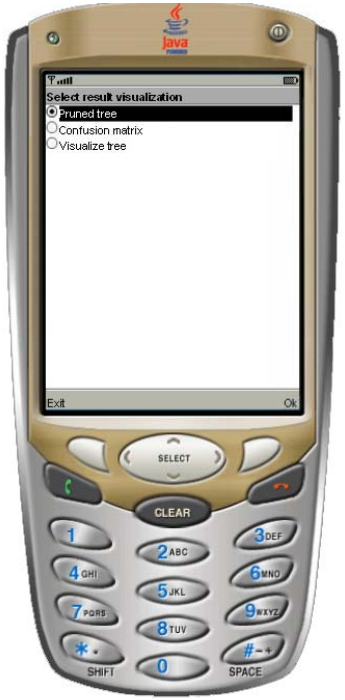
end

schema: “/../DMService.wsdl”
operation: “classify”
argument: “J48”
argument: “-C 0.25 -M 2”
...

schema: “/../DMService.wsdl”
operation: “clusterize”
argument: “EM”
argument: “-I 100 -N -1 -S 90”
...

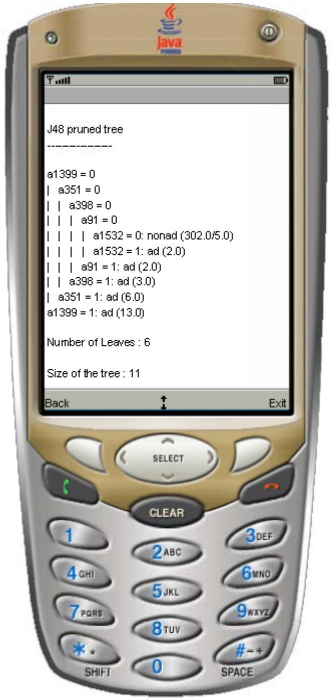
Grid Services for Mobile Data Mining

- A user can choose the mining algorithm and select which part of a result (data mining model) he wants to visualize.



Select result visualization

● Pruned tree
○ Confusion matrix
○ Visualize tree



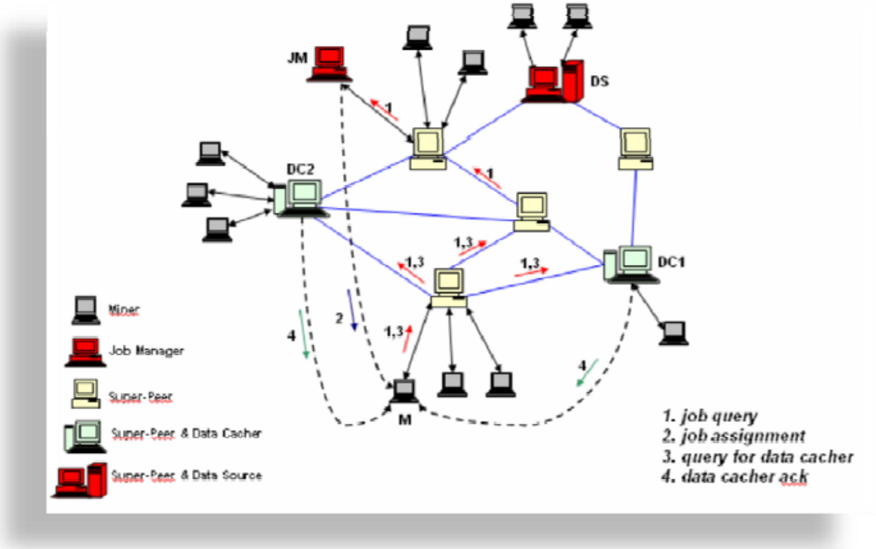
```
J48 pruned tree
------------

a1399 = 0
|   a351 = 0
|   |   a398 = 0
|   |   |   a91 = 0
|   |   |   |   a1532 = 0: nonad (302.0/5.0)
|   |   |   |   a1532 = 1: ad (2.0)
|   |   |   a91 = 1: ad (2.0)
|   |   a398 = 1: ad (3.0)
|   a351 = 1: ad (6.0)
a1399 = 1: ad (13.0)

Number of Leaves : 6

Size of the tree : 11
```

Mining@home

- The Public Resource Computing paradigm (PRC) is currently used to execute large scientific applications with the help of private computers (Seti@home, Climate@home, Einstein@home).

- PRC model can be exploited to program to P2P data mining tasks involving hundreds or thousands of nodes.

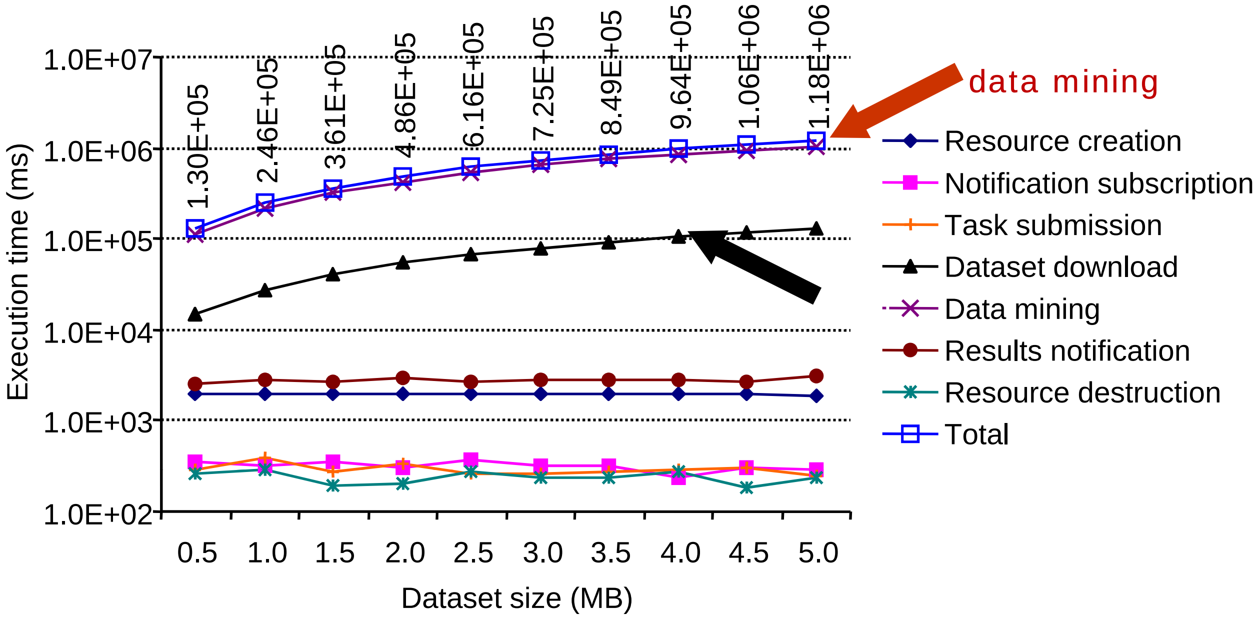
- Highly decentralized data analysis tasks can be programmed as large collections of threads or services.



1. job query
2. job assignment
3. query for data cacher
4. data cacher ack

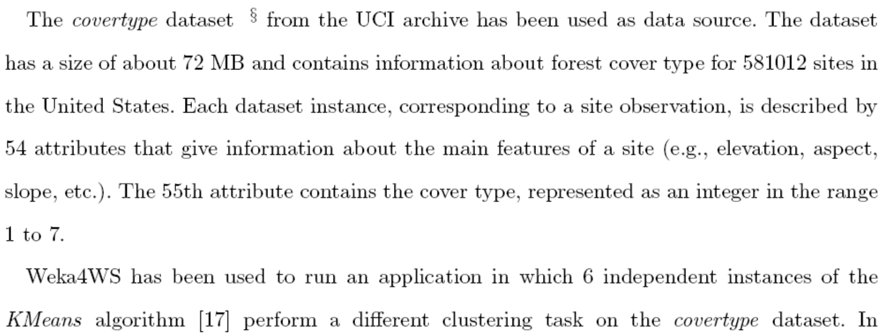
Impact of the WSRF overhead

Execution times



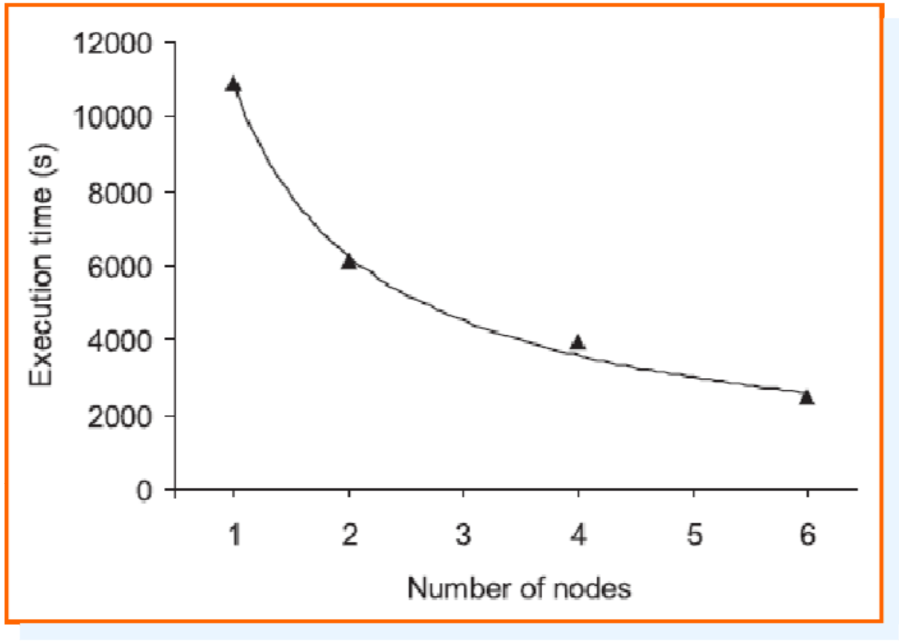
- In a Grid scenario the **data mining step represents from 85% to 88%** of the total execution time, the dataset download takes about 11%, while **the other steps range from 0.5% to 4%.**

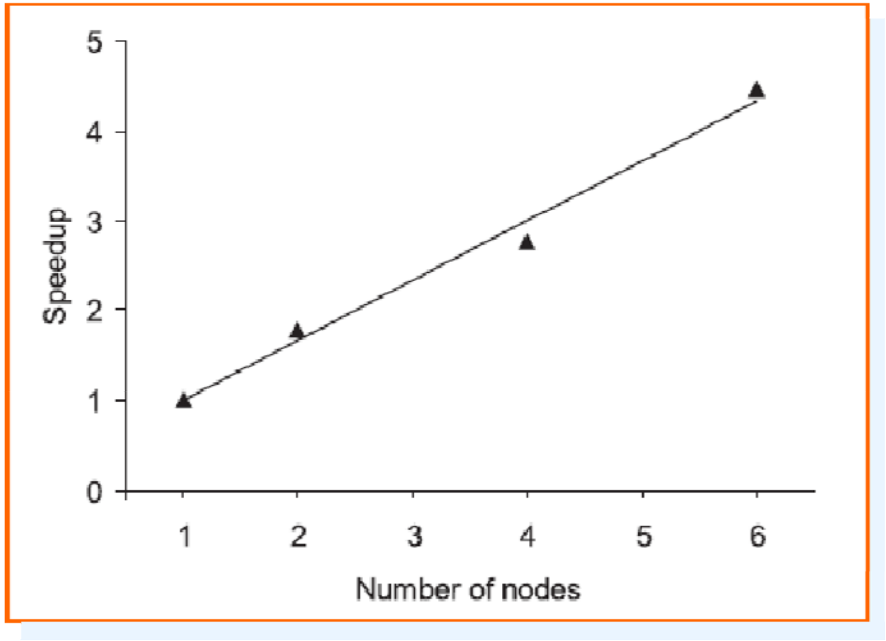
Weka4WS: application speedup on a Grid



The *covertype* dataset [§] from the UCI archive has been used as data source. The dataset has a size of about 72 MB and contains information about forest cover type for 581012 sites in the United States. Each dataset instance, corresponding to a site observation, is described by 54 attributes that give information about the main features of a site (e.g., elevation, aspect, slope, etc.). The 55th attribute contains the cover type, represented as an integer in the range 1 to 7.

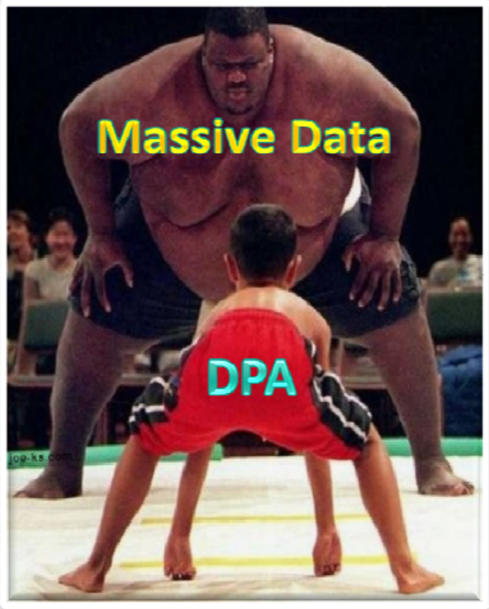
Weka4WS has been used to run an application in which 6 independent instances of the *KMeans* algorithm [17] perform a different clustering task on the *covertype* dataset. In





Summary

- New HPC infrastructures allow us to attack new problems, BUT require to solve more challenging problems.
- New programming models and environments are required
  - Data is becoming a BIG player, programming data analysis applications and services is a must.
  - New ways to efficiently compose different models and paradigms are needed.
  - Relationships between different programming levels must be addressed.



- In a long-term vision, pervasive collections of data analysis services and applications must be accessed and used as public utilities.
- We must be ready for managing with this scenario.

Thanks