Goals

• Introduce some **ongoing research projects** on Cloud Computing.

• Present how data mining applications and KDD processes can be developed as **collections of Cloud analytics services**.

• Outline a Cloud data analytics framework.

• Present **some promising research directions** on multi-agents on Clouds.
Outline

1. Introduction
2. Some research projects on Clouds
3. Big Data & KDD
4. Data Mining Applications as Cloud Services
5. Software Agents and Cloud Platforms
6. Final Remarks
Current Research Projects

• Reliable and Scalable MapReduce
  • A decentralised MapReduce implementation for managing master and slave failures in large scale applications.

• Energy-efficient Data Centers
  • A decentralized self-organizing strategy for energy saving Data Centers.

• Cloud Data Mining Services
  • Service-oriented workflow systems for data mining and KDD applications on Cloud platforms.

• Cloud services for mobile devices/users
  • Virtualization of mobile environments on Clouds.
• Bigger and more complex problems must be solved by large scale distributed computing.

• DATA SOURCES are larger and larger and ubiquitous (Web, sensor networks, mobile devices, telescopes, ...).
...and Big Data

- Even where accessible, much data in many fields cannot be read by humans

  so

- The huge amount of data available today requires smart data analysis techniques to aid people to deal with it

  and

- Scalable algorithms, techniques, and systems are needed.
Data Availability or Data Deluge?

- Today the information stored in digital data archives is enormous and its size is still growing very rapidly.
Some decades ago the main problem was the shortage of information, now the challenge is:

- 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.
Data: From Storing to Analysis

• Storing Data is not the main problem.

• A key issue is analyse, mine, and process Data for making it useful.
Distributed Data Intensive Apps

- In this scenario Cloud computing systems provide an effective **computational and data access support**.

- Cloud (HPC systems and Grids) allows for running **distributed data intensive applications** and **data mining** in large and distributed data sets.

- Clouds can be **used in integrated frameworks** through service interfaces for manage large data sources and process them.
Service-Oriented Data Mining

- Knowledge discovery (KDD) and data mining (DM) are:
  - **Compute- and data-intensive processes/tasks**
  - **Often based on distribution of data, algorithms, and users.**

- Large scale service-oriented systems (like Clouds and Grids) integrate both distributed computing and parallel computing, thus they are key infrastructures for high-performance distributed knowledge discovery. (e.g., Knowledge Grids, Data Analytics Clouds)

- They also offer
  - **security, resource information, data access and management, communication, scheduling, fault detection, ...**
Services for Distributed Data Mining

• By exploiting the SOA model it is possible to define **basic services for supporting distributed data mining tasks/applications.**

• Those services can address all the aspects of data mining and in knowledge discovery processes
  • **data selection and transport services,**
  • **data analysis services,**
  • **knowledge models representation services,** and
  • **knowledge visualization services.**
Collection of Services for Distributed Data Mining

- It is possible to design services corresponding to:

<table>
<thead>
<tr>
<th>Data Mining Applications and KDD processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>This level includes the previous tasks and patterns composed in a multi-step workflow.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distributed Data Mining Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>This level implements, as services, patterns such as collective learning, parallel classification and meta-learning models.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Single Data Mining Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Here are included tasks such as classification, clustering, and Association rules discovery.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Single KDD Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>All steps that compose a KDD process such as preprocessing, filtering, and visualization are expressed as services.</td>
</tr>
</tbody>
</table>
Distributed Data Mining Services

- This collection of data mining services implements an Open Service Framework for Distributed Data Mining.

- Distributed Data Mining patterns
- Distributed Data Mining Applications and KDD processes
- Data Mining Task Services
- KDD Step Services

Open Service Framework for Distributed Data Mining
Distributed Data Mining Services

- It allows developers to program distributed KDD processes as a **composition of single and/or aggregated services** available over a service-oriented infrastructure.

A Service-oriented Cloud workflow

- Those services should exploit other basic Cloud/Web services for data transfer, replica management, data integration and querying.
Data mining services

- By exploiting the Cloud services features it is possible to develop data mining services accessible every time and everywhere (remotely and from small devices).

- This approach can result in
  - Service-based distributed data mining applications
  - Data mining services for communities/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 Services

- Service-based KDD systems we developed
  - Weka4WS
  - KNOWLEDGE GRID
  - Data Mining Cloud Framework
Data Mining Cloud

- **PaaS** (Platform is a Service) is the appropriate model to build a framework that allow users to design and execute data mining applications.

- We used this approach for implementing a prototype of the DM Cloud App and of the Data Mining Cloud Framework on the Windows Azure Cloud.
Data Mining Cloud App

Task Submission

Select algorithm and parameters:

Algorithm:

- Select algorithm ->

- Select algorithm ->
  - weka.classifiers.rules.JRip
  - weka.classifiers.rules.NB
  - weka.classifiers.trees.ADTrees
  - weka.classifiers.trees.kNN
  - weka.classifiers.trees.LB
  - weka.classifiers.trees.NB
  - weka.classifiers.trees.BM
  - weka.classifiers.SimpleKMeans

http://clsmining.cloudapp.net/TaskSubmission.aspx
Data Mining Cloud App

DATA MINING CLOUD APP

Home  Task submission  Task status  About

TASK SUBMISSION
Select algorithm and parameters:
Algorithm: weka.classifiers.trees.J48
Dataset: Select input
- Value of a column:
  - US Census143n.arff
  - US Census286n.arff
  - US Census529n.arff
  - Correlation 0.96MB.arff
  - Correlation 0.95MB.arff
  - Correlation 0.94MB.arff
  - iris.arff
  - weather.arff

Execute

http://datamining.cloudapp.net/TaskSubmission.aspx
Data Mining Cloud App

Task Status:
- Correctly Classified Instances: 146, 97.3333%
- Incorrectly Classified Instances: 4, 2.6667%
- Kappa statistic: 0.96
- Mean absolute error: 0.0293
- Root mean squared error: 0.1209
- Relative absolute error: 6.3821%
- Root relative squared error: 25.0145%
- Total Number of Instances: 150

Confusion Matrix:

```
       a   b   c
a  10  0  0
b  0  47  0
c  0  3  46
```

 Stratified cross-validation:
- Correctly Classified Instances: 143, 95.3333%
- Incorrectly Classified Instances: 7, 4.6667%
- Kappa statistic: 0.91
- Mean absolute error: 0.0429
- Root mean squared error: 0.1742
- Relative absolute error: 4.8752%
- Root relative squared error: 36.9443%
- Total Number of Instances: 150
Data Mining Cloud Framework
• **Data Mining Cloud Framework** for workflow-based KDD applications, expressed as a graphs that link together data sources, data mining algorithms, and visualization tools.
Some Preliminary Experiments

![Graph showing execution time vs. number of nodes for different data sizes (9MB, 18MB, 36MB). The graph indicates that as the number of nodes increases, the execution time decreases. Specific points marked at 13 hours and 2 hours for different data sizes.](image)
Cloud & Agents: Marriage of Interest?

TWO VIABLE APPROACHES

1. **Clouds Using Agents.** Use of agents and their features (flexibility, intelligence, pro-activity, and autonomy) to implement new solutions and services not already available on Clouds.

2. **Agents Using Clouds.** Use of Cloud systems as hardware-software platforms to run large-scale simulations and multi-agent applications at reduced costs.
Clouds Using Agents

• Most attention has been devoted to the production and deployment of Cloud infrastructures to support virtualization and data centers.

• Little attention to introduce innovative techniques for users and developers to discover, obtain, manage, and use Cloud resources and services.

• Autonomous agents and MAS are flexible and smart tools to negotiate user access, automate resource and service discovery and composition, purchase and use Cloud machines.
Clouds Using Agents

- A new research area, **agent-based Cloud computing** will produce agent-based software solutions for Cloud systems:
  - resources and service management and discovery,
  - SLA negotiation, and
  - service composition.

- **Agents can make Clouds smarter** in providing services to users and more efficient in processing allocation and in storage management for Cloud applications.
Clouds Using Agents

- By implementing agent-based solutions in software systems that manage Cloud infrastructures we can have:
  - Cloud services that are more intelligent and flexible,
  - Dynamics and pro-active services,
  - Autonomic Clouds.
Final Remarks (1)

- Scientific and industrial applications must be able to analyze very large data sources (archives, databases, flat files).

- Data mining and knowledge discovery tools are necessary to find what is interesting in them.

- Cloud systems can effectively be used as distributed infrastructures for service-oriented data mining frameworks and applications.
Final Remarks (2)

• Other research issues, such as
  • Techniques for energy efficient data centers,
  • Reliable MapReduce programming for large-scale Clouds,
  • Integration between Clouds and Mobile devices,

are promising and effective for new Cloud solutions.

• New distributed infrastructures allow us to attack new problems, BUT require to solve more challenging problems. Research must be ready to provide new solutions.
QUESTIONS?

Credits:
Eugenio Cesario
Marco Lackovic
Fabrizio Marozzo
Paolo Trunfio