Scalable Data Analytics
Cloud Solutions for Large-Scale Data Analysis

Domenico Talia
talia@dimes.unical.it

www.ScalableDataAnalytics.com
Huge amounts of data are stored every day in the Cloud.

2016: >50% of large companies data will be on the Cloud  

This Big Data must be analyzed in a cost- and time-effective way to make it useful.
1,100,000 companies use public Clouds

66% use Amazon AWS, Google App Engine, or Windows Azure

50% are interested to analytics services

Big Data: $34B in 2013, +21% over 2012
**SDA** is start-up that designed a software system for developing **scalable** data analysis workflows on Clouds

**Key features**

- ✔ Software as a Service
- ✔ Visual and script-based
- ✔ Open to 3rd-party tools
- ✔ Parallel execution
Some applications

**Finance:** Prediction of personal income based on census data

**E-Health:** Disease classification based on gene analysis

**ICT:** Discovery of network attacks from connection log analysis
Cloud Architecture

The system automatically parallelizes workflow execution on multiple virtual machines.

Cloud scalability is exploited for data storage and virtual server allocations.
### Performance and costs

#### Case study*: Log analysis for intrusion detection, 46M tuples

<table>
<thead>
<tr>
<th>Cloud setup (Windows Azure)</th>
<th>Exec. time (hh:mm)</th>
<th>Total cost</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential (1 VM)</td>
<td>107:42</td>
<td>7 $</td>
<td>-</td>
</tr>
<tr>
<td>Parallel (64 VMs)</td>
<td>2:07</td>
<td>9 $</td>
<td>50.78</td>
</tr>
</tbody>
</table>

+2 $ cost  
\( \times 50 \) speedup

*presented at:* Supercomputing’s WORKS13, November 17, 2013, Denver (USA)

### Equivalent non-Cloud alternative**: 86,000 $

**SGI Altix XE, 64-core cluster, 64GB RAM, 2TB disk
Domenico Talia  
Professor of Distributed Systems, UNICAL  
Director of ICAR-CNR

Fabrizio Marozzo  
PhD in Systems and Computer Engineering, UNICAL

Eugenio Cesario  
PhD, Researcher in Computer Science and Information Engineering, ICAR-CNR

Paolo Trunfio  
PhD, Researcher in Information Processing Systems, UNICAL

1st prize at StartCup Calabria 2013

Intel special prize at the Italian National Award for Innovation (PNI 2013)

Microsoft BizSpark Platinum Award at the Italian National Award for Innovation (PNI 2013)