Information Systems

*Introduction to concepts, requirements, approaches, and best-practices for designing Information systems in hybrid data infrastructure*

Pasquale Pagano
• Education
  • Master Degree in Computer Science
  • Ph.D in Information Engineering on Distributed Systems

• Organization
  • CNR – ISTI, InfraScience Group

• Experience
  • D4Science Hybrid Data Infrastructure, Technical Director
  • gCube Open-Source Framework, Technical Director
  • BlueBRIDGE EU Project, Technical Director
  • SoBigData EU Project, Infrastructure Manager
  • Parthenos EU Project, Infrastructure Operation Manager

• Bio and contact
  • it.linkedin.com/in/pasqualepagano/
  • pasquale.pagano@isti.cnr.it

Pasquale Pagano
Outline

Information System
• What it is and how to define it

Context
• Hybrid cloud-based infrastructure

Resource Registry
• Hybrid cloud-based infrastructure information system

Conclusions
Information Systems

An information system (IS) is

- any organized system for the collection, organization, storage and communication of information

- an integrated set of components for collecting, storing, and processing data and for providing information, knowledge, and digital products [Encyclopaedia Britannica]

Information consists of data that is

1. **accurate** and **timely**,  
2. specific and **organized for a purpose**,  
3. presented **within a context** that gives it meaning and relevance,  
4. can increase understanding and **decrease uncertainty**
An information system (IS) is

- a combination of hardware, software, infrastructure and trained personnel organized to facilitate planning, control, coordination, and decision making in an organization [businessdictionary]

Trained personnel consists of human resources and:

1. procedures for using, operating, and maintaining the information system
2. set of basic principles and associated guidelines, a.k.a policies, formulated and enforced to direct and limit actions in pursuit of long-term goals
An information system (IS) is

- a software system to capture, transmit, store, retrieve, and manipulate data produced by software systems to provide access to information, thereby supporting people, organizations, or other software systems [MIT Press]

Software systems become producer and consumer of the Information System making it at the core of their business activities.
Information Systems Definition

A software system

• to capture, transmit, store, retrieve, and manipulate data produced by software systems

• to provide access to information, organized for a purpose and within a contextual domain
  • used, accessed, and maintained according to well-known procedures operated under the limit of the (evolving) organization policies

• to support people within an organization and other software systems
Hybrid cloud-based infrastructure

CONTEXT
e-Infrastructures

e-Infrastructures enable researchers in different locations across the world to collaborate in the context of their home institutions or in national or multinational scientific initiatives. They can work together by having shared access to unique or distributed scientific facilities (including data, instruments, computing and communications).
e-Infrastructures
e-Infrastructures

Data e-Infrastructure: an e-Infrastructure promoting data sharing and consumption. Addresses the needs of the research activity performed by a certain community.
e-Infrastructures

Computational e-Infrastructure: an e-Infrastructures offering computational resources distributed in a network environment. Uses Cloud computing to execute calculations with a large number of connected computers. Offers collaboration facilities for scientists to share experimental results.
Requirements for e-Infrastructures

- Support collaborative research and experimentation
- Implement Reproducibility-Repeatability-Reusability
- Allow sharing of data, methods, workflows, and findings
- Grant open access to produced scientific knowledge and data
- Tackle simplified access to existing computing and storage resources
- Ensure low operational and maintenance costs
- Manage heterogeneous data and service access policies
Virtual Research Environment

An operational environment

- Where set of resources (data, services, computational, and storage resources)

- are assigned to group of users via interfaces

- for a limited timeframe

Created on demand

No cost for the resource providers

Open to host and operate custom software

Regulated by tailored policies

D4Science is both a Data and a Computational e-Infrastructure that federates other e-Infrastructures across administration domains - **Hybrid Data Infrastructure**

Moreover, it
- Implements the notion of e-Infrastructure/platform/software as-a-Service
  - it offers on demand access to data management services and computational facilities;
- is policies-driven through the true implementation of **Virtual Research Environments**
Infrastructure as a Service

Infrastructure as a service (IaaS) is a standardized, highly automated offering, where compute resources, complemented by storage and networking capabilities are owned and hosted by a service provider and offered to customers on-demand.

- IaaS also hosts users' applications and handles tasks including system maintenance, backup, and recovery planning.

- Customers are able to self-provision this infrastructure, using a Web-based graphical user interface that serves as an IT operations management console for the overall environment.

- API access to the infrastructure may also be offered as an option.
Cloud Computing

- IaaS is one of three main categories of cloud computing services, complemented by

- **Software as a Service (SaaS)**
  - software distribution model in which applications are made available to customers over the Internet.
  - removes the need to install and run applications on owned data center.
  - eliminates the expense of hardware acquisition, provisioning and maintenance, as well as software licensing, installation and support.

- **Platform as a Service (PaaS)**
  - cloud computing model that delivers application development frameworks to its users as a service.
Cloud Computing Characteristics

• On-demand
  • Provision of computing resources, such as server, service, and storage, as needed without requiring human interaction

• Broad network access
  • Resources are available over a network

• Resource pooling
  • Resources pooled to serve multiple users using a multi-tenant model, with physical and virtual resources dynamically assigned and reassigned according to consumer demand

• Rapid elasticity
  • Resources elastically provisioned and released, automatically, to horizontally scale rapidly outward and inward as needed

• Measured service
  • Resources usage is monitored, controlled, and reported
D4Science is an **hybrid cloud-based infrastructure** technologies integrated to provide elastic access and usage of data and data-management capabilities

**Humanities and Cultural Heritage**

**Social Mining**

**Environmental Studies**

**Biological and Ecological Studies**
D4Science Service Provision

Empowered Hardware

Package Repository

Dynamic deployment

Production

Package Repository

Rapid deployment

Failure Recovery

Service provision continuity

Dynamic Load Balancing

Balancing utilization with head room

Context
D4Science is an hybrid cloud-based infrastructure

- **63 VREs** hosted
- **+3100 users**
  - in **44 countries**
  - from **+80 Institutions**
- **+ 430 millions service calls a year**
- **+ 1600 distinct caller hosts**
- **+25,000 derivative data/month**
- **+50 data providers**
- **over a billion quality records**
- **+20,000 temporal datasets**
- **+50,000 spatial datasets**
- **99.8% service availability**
Hybrid cloud-based infrastructure challenges

Hundred software systems opportunistically deployed on demand

- The software systems to manage are not known at design time
- The location of any service is known only at runtime
- Any software system has to discover the location of the targeted service before to use it
- All software systems have to be monitored, controlled, and reported
- Status, load, exploitation usage, and accounting data have to be constantly updated to enable elasticity and pooling of resources

All these data are managed by the infrastructure Resource Registry
Hybrid cloud-based infrastructure information system

RESOURCE REGISTRY
Resource Registry

The infrastructure Resource Registry is an Information System designed to support the operation of an hybrid cloud-based infrastructure

• To capture, transmit, store, retrieve and manipulate data from any software system enabled on the infrastructure
  • Location and properties
  • Status, load, exploitation usage, and accounting data

• To provide access to information, organized to enable
  • Monitoring, validation, and reporting
  • Elasticity and pooling of resources

• To support any software system to
  • Discover services and infrastructure resources
The Resource Registry - core of a SOA within the complexities of an hybrid cloud-based infrastructure – must enable

- a set of resource management functions
  - enabling functions
    - publication, discovery
    - monitoring, deployment
    - contextualization, security, execution
  - data management functions
    - access, store
    - index, search
    - transfer, transform

- plus a set of applications
  - built against those functions
Resource Registry

abstract system view

- **Resource types: abstract view over functions**
  - defined by specifications
  - multiple implementations, over time / concurrently

- **different implementations, different information**
  - system cannot globally define them
  - implementations produce/consume different *facets*, independently

- **resource semantics dynamic**
  - no longer predefined in class hierarchies
  - implicitly captured by current facets
  - changes over time / across “similar” resources
Resource Registry
Resource Registry

resource model

- defines a framework for collecting facets
  - some common properties
  - a loose binding to XML/Json

- all resources have:
  - A unique identifier
  - optional name and description
  - one or more policies
  - zero or more facets
    - uniquely identified
    - arbitrary otherwise
Resource Registry

resource model
Resource Model

Entities and Relations

- **Resource**
  - **Resource Instance**
    - **Resource Schema**
      - **Context**
        - **Application Domain**
          - **Facet Instance**
            - **Facet Schema**

Relations:
- **isDescribedBy**: 1..1
- **belongsTo**: 1..n
- **isRelatedTo**: 0..n
- **contains**: 1..n
- **defines**: 1..n
- **consistsOf**: 1..n
Resource Model

Entities and Relations
Resource Model

milestones

• Open-ended model for describing resources

• Open-ended set of manageable resources

• Ability to evolve with the evolving needs of the infrastructure at no cost for its clients
  • by supporting new types of resources at run-time
  • by supporting evolution in the way a resource is described
  • by supporting the same resource type described by using different models
Resource Registry architecture

Any Service

Resource Registry Client

Resource Registry Publisher

High Availability Proxy

Context Management
Schema Management
Entity Management
Query & Access

Context Management
Schema Management
Entity Management
Query & Access

IS-Model

Graph DB

Graph DB

Graph DB
Conclusions

• Any information system has to be designed for a purpose and within a contextual domain

• A Resource Registry is an Information System designed to support the operation of an infrastructure
  • Open-ended model since infrastructure resources may not known in advance
  • Open-ended set of manageable resources since an infrastructure lifetime may span several decades
  • Non-functional requirements - e.g. availability, reliability – are key requirements to consider in the design phase
Further Reading

• Candela, Leonardo, Donatella Castelli, and Pasquale Pagano. "Virtual research environments: an overview and a research agenda." Data Science Journal 12.0 (2013): 65-91
THANK YOU

Acknowledgement:
Fabio Simeoni, Luca Frosini, Manuele Simi
CNR – ISTI InfraScience

The content of this presentation is released under the
Creative-Commons CC-BY-SA license