DataCloud
Project Structure

Version 1.5

11/7/2014

The Context of the EINFRA-1-2014 Call

- **EINFRA-1-2014**, “Managing, preserving and computing with big research data”
  - Research and Innovation Action (RIA)
- **What**: “Development and deployment of integrated, secure, permanent, on-demand service-driven, privacy-compliant and sustainable e-infrastructures incorporating advanced computing resources and software.”
- **Why**: “increase the capacity to manage, store and analyze extremely large, heterogeneous and complex datasets[1], including text mining of large corpora.”
- **How**: “[P]rovide services cutting across a wide-range of scientific communities and addressing a diversity of computational requirements, legal constraints and requirements, system and service architectures, formats, types, vocabularies and legacy practices of scientific communities that generate, analyse and use the data.”

[1]: Research data include large datasets collected, developed or generated for/by research, integration of small distributed datasets, as well as data not originally collected for research, which may include environmental, social and humanities data.

Some EINFRA-1-2014 Facts

- Budget for EINFRA-1-2014: 55 M€
- Key **excellence criteria** that *will* be considered by the EC:
  - The extent to which the **Networking Activities** will foster a culture of co-operation between the participants and other relevant stakeholders.
  - The extent to which the **Service activities** will offer access to state-of-the-art infrastructures, high quality services, and will enable users to conduct excellent research.
  - The extent to which the **Joint Research Activities** will contribute to quantitative and qualitative improvements of the services provided by the infrastructures.
- Timeline:
  - Deadline: 2/9/2014 at 17:00 Brussels time.
  - Outcome of the evaluation (**single stage**): maximum 5 months from the final data for submission.
  - Indicative date for the signing of grant agreements: maximum 3 months from informing applicants they have been successful.
- Details are in http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-infrastructures_en.pdf

Topics in EINFRA-1-2014

1. Establishing a federated pan-European data e-infrastructure.
2. Services to ensure the quality and reliability of the e-infrastructure.
3. Federating institutional and, if possible, private data management and curation tools and services.
4. **Large scale virtualization of data/compute centre resources.**
5. **Development and adoption of a standards-based computing platform (with open software stack).**
6. Support to the evolution of EGI.
7. Proof of concept and prototypes of data infrastructure-enabling software (e.g. for databases and data mining) for extremely large or highly heterogeneous data sets.
8. Enable the creation of a platform and infrastructure for mining text aggregated from different sources/publishers.

EINFRA-1-2014: Items 4-5

- **Item 4:**
  - Large scale virtualization of data/compute centre resources to achieve on-demand compute capacities, improve flexibility for data analysis and avoid unnecessary costly large data transfers.
- **Item 5:**
  - Development and adoption of a standards-based computing platform (with open software stack) that can be deployed on different hardware and e-infrastructures (such as clouds providing infrastructure-as-a-service (IaaS), HPC, grid infrastructures…) to abstract application development and execution from available (possibly remote) computing systems. This platform should be capable of federating multiple commercial and/or public cloud resources or services and deliver Platform-as-a-Service (PaaS) adapted to the scientific community with a short learning curve. Adequate coordination and interoperability with existing e-infrastructures (including GÉANT, EGI, PRACE and others) is recommended.

DataCloud

- The key goal of the project is to respond to topics 4 and 5 of the EINFRA-1-2014 call:
  - Bringing together European excellence and experience in developing solutions for scientific distributed computing.
  - Providing an overall solution *enabling* diverse scientific communities to expand and improve their activities.
  - Testing the developed solution within key private and public e-Infrastructures and in collaboration with large European resource providers.
  - Aligning development to compelling use cases provided by significant scientific communities.

DataCloud Technology / Resource / Infrastructure Providers / Industries

- **INFN** (IT)
- **EGI.eu** (NL)
- **CSIC** (ES)
- **CERN** (CH)
- **KIT** (DE)
- **DESY** (DE)
- **LIP** (PT)
- **IN2P3** (FR)
- **CESNET** (CZ)
- **PSNC** (PL)
- **Cyfronet** (PL)
- **CEA** (FR)
- **CMCC** (IT)
- **UPV** (ES)
- **T-Systems International GmbH**
- **Atos S.E.**
- **Santer Reply S.p.A.**
- **<to be confirmed>**

DataCloud User Communities

- **LBT**, through INAF
- **CTA**, through INAF
- **LifeWatch**, through CSIC
- **WeNMR/INSTRUCT**, through University of Utrecht and University of Florence
- **EMSO**, either directly or through INGV
- **BBMRI**, through UPV
- **EuroBioImaging**, through UPV
- **DCH-RP**, through ICCU
- **DARIAH**, either directly or through a DARIAH institution
- **ENES**, through CMCC
- **WLCG**
- **ELIXIR-IT**, through CNR
- **ELIXIR-CZ**, through CESNET
- <Letters of Support under discussion>

WP Structure

- WP1: Administrative and Technical Management
- WP2: Requirements, outreach, collaborations and sustainability
- WP3: Product evolution, quality management and application adaptation
- WP4: Resource Virtualization
  - 4.1: Cloud Computing Virtualization
  - 4.2: Cloud Storage Virtualization
  - 4.3: Network Virtualization
- WP5: PaaS Platform
  - 5.1: PaaS architecture
  - 5.2: Security/AAI
  - 5.3: High-level geographical scheduling
  - 5.4: Data access and management
- WP6: Portals, Big Data Analysis and Scientific Workflows
  - 6.1: API's and Toolkits
  - 6.2: Portals and User Interfaces
  - 6.2: Support for big data workflows for eScience

WP1 (NA) Administrative and technical Management

- **Overall administrative management**
- **Management of costs, payment transfers, and of partner resp. WP reports**
- **Work package coordination**
- **Management of the project deliverables**
- **SLA and license policy management**
- **Definition of the project technical management plan**

WP2 (NA) Requirements, outreach, collaborations and sustainability

- **2.1 Requirement gatherings from participating communities**
  - We will start from real use cases provided by participating communities. There will a set of selected apps, representative of these use cases (app selection will be based on several criteria, e.g. size of a community, impact, sustainability, number of present and prospective users). The apps themselves should be concisely described in the proposal.
  - CSIC, UPV, User Communities, EGI.eu
- **2.2 Knowledge Management**
  - (to be clarified)
  - EMSO
- **2.3 Dissemination and Public Relations**
  - This may include identification and possibly support for “external” applications, i.e. coming from communities external to the project.
  - EGI.eu, DARIAH, EMSO
- **2.4 Sustainability and Exploitation management**
  - EGI.eu

WP3 (SA) Product evolution, quality management and application adaptation

- **3.1 Pilot infrastructure management**
  - Note: we need to have two infrastructures:
    - Integration infrastructure, for evaluation of development and interoperability.
    - Production-level infrastructure, for application deployment. This is a subset of what may already exist in production and be built on top of existing DCIs (e.g. EGI Grid, EGI FedCloud, HPC centers).
  - DESY, IN2P3, INFN, CESNET, CEA, CSIC, LIP, INAF, CMCC, CERN
- **3.2 Adaptation and validation of selected applications**
  - The selected applications should be deployed on the "production-level" infrastructure to validate the development in a real environment.
  - CSIC, UPV, User Communities, EGI.eu
- **3.3 Quality assurance policies and enforcement.**
  - Note: how do we handle the software quality problem? (uptime? Bug tracking? Identification of KPIs? E.g. feedback on number of fulfilled user requirements.)
  - LIP, EGI.eu
- **3.4 Product release policies and management**
  - LIP, INFN
- **3.5 Software and Product maintenance**
  - INFN

WP4 (JRA) Resource Virtualization development

- **4.1 Cloud Computing Virtualization**
  - Note: contextualization of containers for application development and deployment, creation of trusted repositories, local scheduling policies for Cloud frameworks.
  - CSIC, KIT, IN2P3, INFN, CESNET, UPV, LIP, CEA, INAF, Atos
- **4.2 Cloud Storage Virtualization**
  - Note: focused on local data centers.
    - To simplify storage management: letting administrators to dynamically provide the storage technology required by end users.
      - One virtual single pool of storage that could export posix, block and object storage in a flexible way.
    - Provide users with a more general interface to access the storage without the need to know about the technology of the available storage
  - DESY, KIT, IN2P3, INFN, Cyfronet, INAF, CMCC
- **4.3 Network Virtualization**
  - Note: analysis of the state of the art with regard to dynamic network virtualization (e.g. OpenNaaS), development of a solution for dynamic and transparent network (hybrid) connections across data centers.
  - DESY, INFN, LIP, INAF

WP5 (JRA) PaaS Platform development

- **5.1 PaaS architecture**
  - Identify an existing open source PaaS framework to be extended by DataCloud and define suitable interfaces to WP4 and WP6. Integrate development occurring in this WP into the PaaS framework.
  - CSIC, INFN, CERN, Atos, T-Systems, Reply
- **5.2 Security and AAI**
  - DataCloud will be agnostic and use whatever kind of AuthN is available and/or provided by other projects. The focus will be on distributed AuthZ.
  - DESY, KIT, IN2P3, CESNET, INFN, CEA, Cyfronet, INAF, CERN
- **5.3 High-level geographical scheduling (APIs)**
  - These APIs should provide a single instance of a computational farm to users, relying also on developments made in WP4. they might also provide advanced features related e.g. to advance reservation, workflow management. We need to put together different technologies such as Grids, Clouds, Desktops, etc.
    - Possible extensions: QoS requests for a given application/job, advanced reservation of resources, dependencies among submitted tasks, decision making engine to determine whether it is better to move data towards the computational resources or not.
  - DESY, CESNET, INFN, UPV, CERN, Atos, T-Systems, Reply
- **5.4 Data access and management**
  - Provide the capabilities to access data remotely, focusing on easy to use and high-level features to move data not only among data centers, but also to/from external data sources (e-infrastructures such as EUDAT, EGI, PRACE, but also simple external services like public web/ftp servers, etc.)
  - Cyfronet, DESY, INFN, INAF

WP6 (JRA) Portals, Big Data Analysis and Scientific Workflows

- **6.1 API's and toolkits**
  - API's that can later be used by portals, desktop and mobile applications.
  - Extend the concept of SAGA adaptors to commercial cloud middleware stacks (EC2, Azure, etc.)
  - Make open storage cloud and commercial (personal) storage clouds interoperable at user application/data level
  - INFN, PSNC, Cyfronet, IN2P3
- **6.2 Portals and User Interfaces**
  - User-friendly portals for selected user communities/scenarios
  - Integration with A&A technologies: SAML-, Oauth-, OpenID- and STORK-based identity federation
  - Integration with Scientific Workflow services
  - Development of toolkit and interfaces for mobile appliances to support selected use cases
  - INFN, PSNC
- **6.3 Support for big data workflows for eScience**
  - Workflow support for big data analytics (Analytics as a Service)
  - Dynamic scientific workflows (Workflows as a Service) providing capabilities for interactive workflows (include the possibility to seamlessly execute workflows on Grid/ Cloud/HPC computing resources with input/output data stored on Grid/Cloud/local storage resources using standards like SAGA, OCCI and CDMI)
  - CMCC, PSNC

WP leaders and deputies

- **WP1**
  - Davide Salomoni (INFN), Isabel Campos (CSIC)
- **WP2**
  - Jesus Marco (LifeWatch), Peter Solagna (EGI.eu)
- **WP3**
  - Jorge Gomes (LIP), Michel Mur (CEA)
- **WP4**
  - Patrick Fuhrmann (DESY), Marcus Hardt (KIT)
- **WP5**
  - Giacinto Donvito (INFN), Lukasz Dutka (Cyfronet)
- **WP6**
  - Marcin Plociennik (PSNC), Roberto Barbera (INFN)