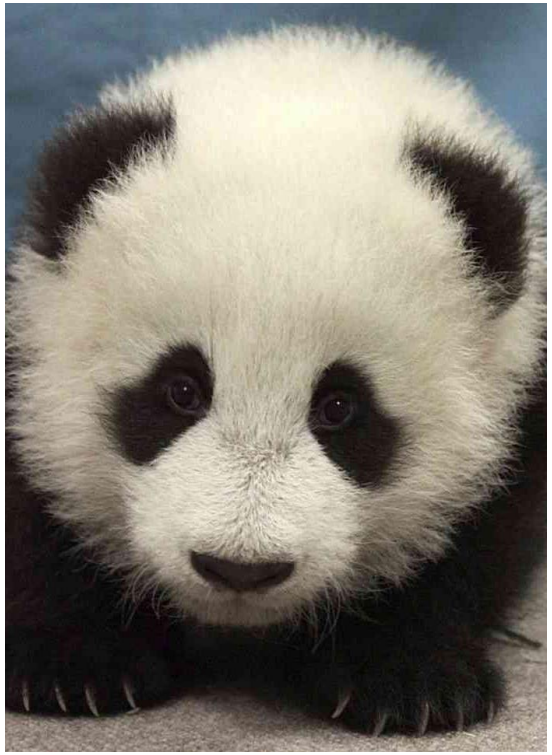


H2020

PaN-DAaS

Big Data Proposal

BIG DATA – ONE BUZZWORD … DIFFERENT CHALLENGES

Industry

- Data mining
- Business intelligence
- Get additional information from (often) already existing data
- Data aggregation
- New field to make money
- Adapted to Hadoop

Science

- **Handling huge amounts of data**
  - Data transport
  - Distributed data sources and/or storage
  - (Global) data management
  - Data preservation
- **Analysing huge amounts of data**
  - Complexity of code
  - Parallel architectures
  - Different software environments
  - Scientific results must be verifiable

BIG DATA IN PHOTON SCIENCE

- **Many research areas, where the data growth is very fast**
  - Cultural heritage, chemistry, earth sciences, life sciences, …
- **Data sets become too big to take home**
- **Data rates require dedicated IT infrastructures to record and store**
- **Data analysis requires farms and clusters. Single PCs not sufficient**
- **Collaborations require distributed infrastructures and networks**
- **Data management a challenge**
- **Software environment a challenge (complexity, heterogenity)**
- **Less IT literate scientists than e.g. in HEP**
- **Users require more support to install software and analyse data**

BIG DATA IN PHOTON SCIENCE

- **PCO-Edge**
  - 2x2kpixels
  - 8 MB images
  - 1000 frames/s
  - 800 MB/s
  - Several TB/day

- Dectris EIGER
  - 2x2k pixels
  - 12 MB images
  - 750 frames/s
  - 1GB/s compressed
  - 3 TB/h

- **Data does not fit any more on USB drives**
- **Users are usually not affiliated to the facility**
- **Users are from many different domains like physics, life-sciences, chemistry, material sciences, cultural heritage …**

PAN-DAAS – MOTIVATION

The big data volumes being generated at photon and neutron sources are increasingly a problem for users because:

(1) they are difficult / impossible to export and
(2) the software to analyse them is not easily available/accessible/usable for many users.

This situation creates an important barrier and means users of these sources are not exploiting the data optimally and new users are not being attracted.

With the unavoidable trend to bigger data volumes it is critical to propose a solution.

New users are increasingly unfamiliar with synchrotron radiation software and need help analysing their data.

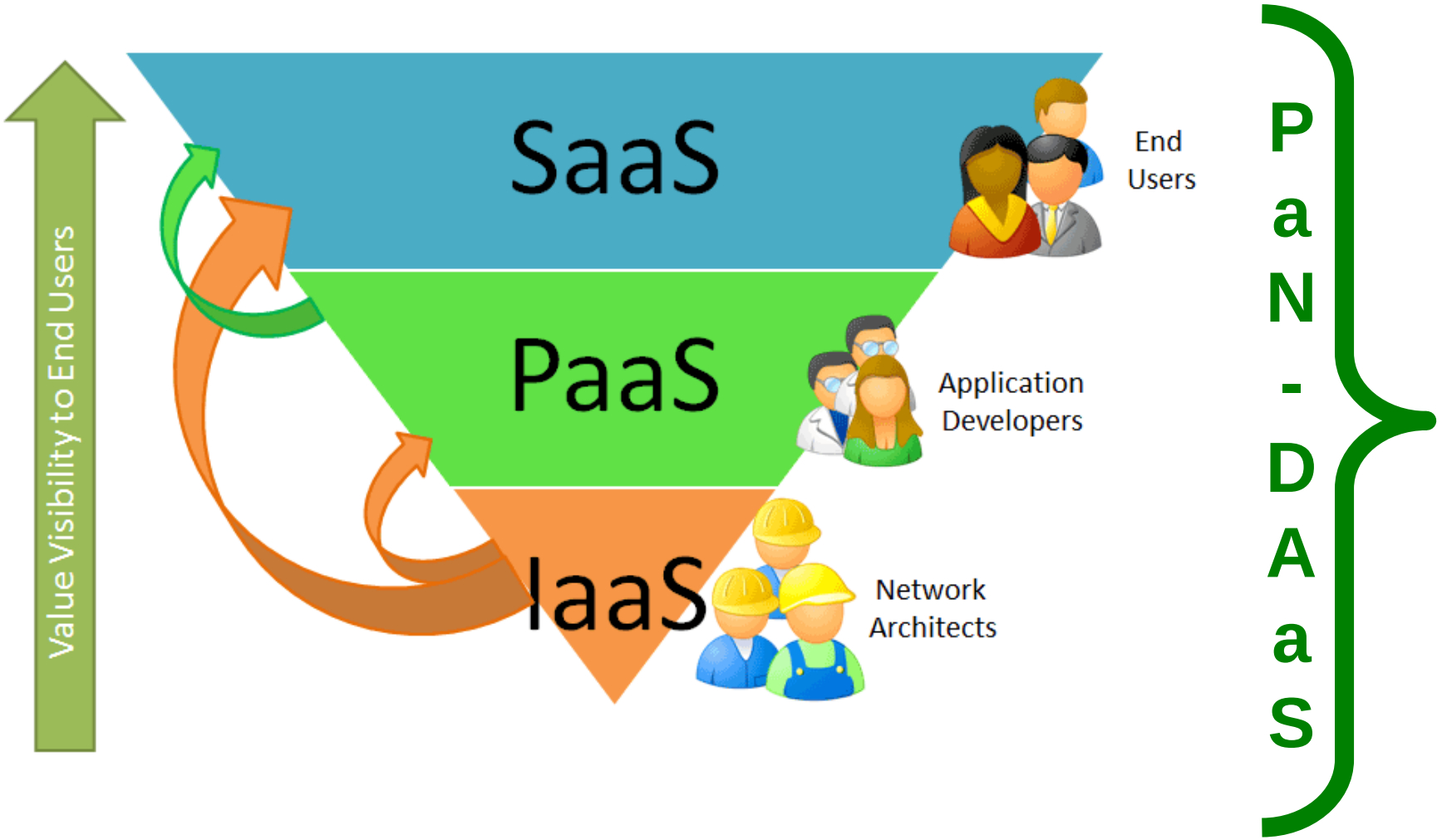
A PARADIGM SHIFT in the approach to DATA ANALYSIS is needed to treat the increasing data volumes created by improved sources, detectors, and software.

PAN-DAAS - MISSION

Mission – The goal of this proposal is to provide users a software and hardware platform to analyse, visualise and browse big data volumes both on site and off site:

- ✓ **The platform must be an all-in-one solution which does not require users to export their data nor install software stacks on their own computer.**
- ✓ **The platform shall be validated with use cases from each of the participating categories of photon and neutron sources.**
- ✓ **Ideally users will be provided with online support to help them analyse their data.**
- ✓ **Solution also applicable to standard techniques with small quantities of data.**
- ✓ **We will not develop new data analysis code but optimise and install existing ones for DAaS.**

PAN-DAAS W.R.T. OTHER SERVICES



PaN-DAaS overlaps over all three Service platforms

H2020 PROGRAMME CHAPTER

ESRF has decided that PaN-DAaS is of highest priority!

Targeting INFRADEV-4 – the ESFRI cluster line (95M€):

Implementation and operation of cross-cutting services and solutions for clusters of ESFRI and other relevant research infrastructure initiatives

- ***Coordinate common activities,***
- ***Define harmonised policies for access to the infrastructures and data lifecycle (acquisition, access, deposit, sharing and re-use),***
- ***Develop and deploy underpinning technologies and services,***
- ***Implement common and efficient solutions for data sharing and provision, architecture of distributed infrastructures, distributed and virtual access management,***
- ***Development of common critical physical and virtual components (e.g. detectors, components for data management).***

H2020 INFRADEV-4 PROPOSAL

Participating ESFRI projects: ILL 20/20, ELI-ALPS, E-XFEL, ESRF-UP

Expressions of interest as of today:

- **PSI, Elettra, Soleil, Diamond, STFC Sci-Comp, Cyl, Sesame**
- **Strong commitment of participating labs needed (+50% investment in hardware and human resources…sustainability!)**
- **Gender balance!**
- **Each participating lab to provide a management statement of support**

Budget: 10-12M€ = ~200 person-years!

Duration: 3-4 year; ESRF proposes 3.5 years (6 months startup phase)

Project Name: Big-PaNDA, PaN-DA, PaNDA, PaN-DAaS, ???

Logo: your imagination is the limit…

PROPOSED PAN-DAAS WORK PACKAGES

- WP1 – Management
- WP2 – Outreach and dissemination
- WP3 – Service architecture
- WP4 – Data browsing + visualisation
- WP5 – Infrastructure platform
- WP6 – Software environment
- WP7 – Use cases 1, 2, 3, and 4
- WP8 – User support – training
- WP9 – Business model, sustainability
- WP10, 11, 12?

WORK PACKAGE 1 – MANAGEMENT

WP1 Objectives

Efficient management of the PAN-DAaS project, in particular carrying out all administrative, legal and financial tasks related to the management of the Grant Agreement; monitor the progress of the work, ensure that appropriate actions are taken. Foster and maintain a collegial team spirit. (~40-45 p-month)

- ✓ **Management of project monitoring**
- ✓ **Communication and interaction with the Commission**
- ✓ **Management of PaN-DAaS internal communication**
- ✓ **Management of meetings**
- ✓ **Management of project reporting**
- ✓ **Production of posters**
- ✓ **Production of short articles for existing periodicals of the RIs**
- ✓ **Steering Committee – implication of users**
- ✓ **User forum**
- ✓ **Industry advisory panel**

WORK PACKAGE 2 – OUTREACH AND DISSEMINATION, INDUSTRY RELATED ACTIVITIES

WP2 Objectives

Dissemination of the PaN-DAaS concepts and the results of the technical work to the PaN-DAaS participants, consortium members, the user communities, present and future industry partners, and science policy makers. Organisation of topical workshops as a forum of exchange of experience. (~24 p-month)

- ✓ **Public web site**
- ✓ **Articles in specialised media**
- ✓ **Participation in events organised by funding bodies with focus on connections to industry**
- ✓ **Organisation of workshops for PaN-DAaS**
- ✓ **Liaise with our user communities**
- ✓ **Liaise with RDA, NERSC**
- ✓ **Organise 3 open conferences**
- ✓ **Organise 3-6 workshops**

WORK PACKAGE 3 – SERVICE ARCHITECTURE

WP3 Objectives

Develop a detailed technical blueprint with an agreed vision to share a common service architecture across RIs. Implement a prototype architecture which is evaluated and used by the scientific community. (~100 p-month)

Evaluate and select an existing architecture

Document on technical blueprint

Select and implement the architecture in all partner RIs

WORK PACKAGE 4 – DATA BROWSING + VISUALISATION

WP4 Objectives

Develop and implement a web-based data browsing and visualisation platform. Data access will be secured with user authentication and authorisation. (~200 p-month)

- ✓ **Authentication using Umbrella**
- ✓ **Authorization based on Umbrella accounts**
- ✓ **User interface platform**
- ✓ **Metadata browsing and searching**
- ✓ **Web based n-dimensional visualisation techniques**
- ✓ **Support for multiple frameworks (e.g. DAWN, MANTID, BASTILLE)**

WORK PACKAGE 5 – INFRASTRUCTURE PLATFORM

WP5 Objectives

Install, commission, and demonstrate with the use cases of WP7 a remotely accessible IT infrastructure for data analysis. (~200-300 p-month)

- ✓ **Evaluation of Best Practices (networks, file-systems, cluster admin, batch schedulers, etc.)**
- ✓ **Evaluation of cloud interfaces (Openstack et al.)**
- ✓ **Latency issues**
- ✓ **Remote access to HPC resources**
- ✓ **Monitoring and accounting**

WORK PACKAGE 6 – SOFTWARE ENVIRONMENT

WP6 Objectives

Select and implement a virtual software environment for running common data analysis and visualisation software. (~300-400 p-months)

- ✓ **On-line and off-line data analysis**
- ✓ **All-in-one software environment**
- ✓ **Web based visualisation software (to be developed)**
- ✓ **Select and implement data analysis software for use cases**
- ✓ **Tutoring interface?**

WORK PACKAGE 7 – USE CASES 1, 2, 3 AND 4

WP7 Objectives

Provide a real service to real users. Use cases for testing service architecture and infrastructure platform throughout the project life. (~36 p-month/participating partner)

Evaluation criteria: must have a scientist and a user community to evaluate service adequacy.

- ✓ **Use case 1: Synchrotrons**

  Life-Sciences, SAXS, tomography
- ✓ **Use case 2: FELs**
- ✓ **Use case 3: Lasers**
- ✓ **Use case 4: Neutrons**

WORK PACKAGE 8 – USER SUPPORT AND TRAINING

WP8 Objectives

Provide on-line user support and assistance for data analysis services. (~100-150 p-month).

- ✓ **Documentation**
- ✓ **Tutoring, MOOC**
- ✓ **Summer school(s)**
- ✓ **Code camps**
- ✓ **User forums**
- ✓ **Social media**

WORK PACKAGE 9 – BUSINESS MODEL AND SUSTAINABILITY

WP9 Objectives

Develop a shared business model guaranteeing long-term sustainability of the DAaS infrastructure. (~1 p-month/participant)

- ✓ **Study long-term funding models**
- ✓ **Study business models of PRACE or other HPC centres**
- ✓ **MoU between partners**
- ✓ **Federation concepts**
- ✓ **Resource allocation models**

WORK PACKAGE 10, 11, 12?

WPxx Objectives

Parallelisation of code on GPU's, Xeon-Phi's?

…your input please!

PROJECT MATRIX

√ = leads √ = participates

RI	WP1 Managemt.	WP2 Outreach	WP3 Service Archi.	WP4 Visualisation	WP5 Infr. Platform	WP6 Soft environ.	WP7 Use cases	WP8 User Support	WP9 Sustainability	WPxx
ESRF	√	√			√	√	√	√	√	
ILL		√					√	√	√	
E-XFEL		√				√	√	√	√	
ELI		√			√		√	√	√	
DESY		√						√	√	
PSI		√						√	√	
ELETTRA		√						√	√	
SOLEIL		√						√	√	
CYI		√						√	√	
SESAME		√						√	√	
STFC-SciC		√						√	√	
DIAMOND		√						√	√	
ISIS		√						√	√	
ALBA		√						√	√	
HZB (?)										
MAX-IV										
Rackspace (?)										
BULL (?)										
KIT										

TIMING

When	What
Now	Constitution of the Consortium
Now	Constitution of an editorial team + collaborative plaform
Now	Nomination of contact person(s) in each RI/company (one IT, one scientist)
Week 15	Letters of support from RI managements
From week 15 onwards	Weekly telco of editorial team
Now till end of June	Proposal writing
02-09-2014 17:00	Proposal submission
Summer next year	Start of the project if accepted

REFERENCES

- **NIST Cloud Computing Definitions**
  - http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf
- **PaaS Best practices**
  - https://www.ibm.com/developerworks/cloud/library/cl-cloudindustry1/
- **ALS's Data Analysis portal**
  - http://spot.nersc.gov/
  - http://cs.lbl.gov/news-media/news/2013/big-data-hits-the-beamline/