Idea for GN4

Purpose: This NIF form is to be used for the submission of New Ideas suggested for inclusion in the GN4 Phase1 and beyond proposals. Budget estimates, information about objectives, impact, benefits, etc. as well as scope must all be supplied.

Submit to: pmo@GÉANT.net by January 31st, 2014 with the subject label starting: GN4Input

Overview

Project Name:	Data Analysis as Service (DaaS)	**Project Proposer:**	Gurvinder Singh, UNINETT

Project Type: GN4 Phase1	Two phase	**Estimated Project Costs (best effort!)**	
Duration proposed	First phase (1 Year): * Study the big data landscape * Demonstrate possibility of the data analysis service using Netflow data. Second phase (2016 -): * Make the service production ready * Implement the service at Inter-NREN scale. * Further research in this fast evolving area.	Manpower in person-months also identifying specific expertise required	10 Months Expertise in Distributed Storage & Computing, Databases, Algorithms, Monitoring
Deliverables proposed (If any can be defined at this stage)	Demo of DaaS applied to data such as Netflow. Deliever a report from study of: * Different distributed data processing systems * Monitoring in such a heterogenous distributed systems * Future directions in this area.	Hardware and equipment:	None

Milestones proposed (If any can be defined at this stage)		Other costs :	15,000 Euro for travel cost

1 Background and Reasoning

Provide background information and the context of the project. Explain the reason for the project. What do you want to be different? What do you hope to improve? Why is the project needed? This should be the reason for the project, not the solution.

In recent years, the amount of research data generated has increased exponentially. So there has been an increasing demand of getting this data to work by storing and processing it in a scalable way. This causes rise of commercially backed open source softwares by the main global actors e.g. Google, Facebook, Twitter, Yahoo. These distributed softwares utilize commodity hardware to store the big data and provide ability to process it locally using data locality paradigm, thus providing good economy of scale.

These systems offers the possibility to store information at scale of 100s of petabyte using distibuted file systems on commodity servers and also offers possibility to process the data in distributed manner e.g. using Map-reduce. The data is replicated across commodity servers to provide releability and scalability. The landscape in this area is evolving very rapidly and systems are evolving to support machine learning algorithms, which is a requierment when analysing such a large amount of data.

Researchers from areas such as Bioinformatics, computer science, astronomy, environment have huge data sets. This project will investigate the possibility of providing a common infrastructure to researchers where they can store and process their data using advanced algorithms e.g. machine learning at a big scale. Individual researchers or institutions from different areas are limited due to resource constraints on how much data they can analyse for their research. Moreover if they have resources available, they has to spend large amount of time in setting up an infrastructure to process their big data and it requires advanced skills at the system level which makes it challenging. So this project will try to investigate the possibility of helping researchers by providing a platform as a service.

In this way, we can contribute in building an eco system where researchers help each other in evolving the service by adding new functionalities and thus improving research further.

2 Objectives, Impact and Benefits

Provide one or more bullet points to briefly describe the primary objective(s) of the project in terms of the desired outcomes. This should be expressed in the form: ‘To ensure…’, ‘To implement…’, 'To service...', 'To improve...', 'To innovate...', 'To optimize...', 'To save...', etc. For each objective mention the benefits to identified stakeholders (e.g. end-users, NRENs, large international research projects, industrial research partners, high level education, etc.) should be mentioned. A description of the expected overall impact must also be provided.

- To innovate in the area of data analysis at a big scale.
- To provide advaced resources to researchers as well as NRENs by giving them a platform to get knowledge out of raw information irrespective of data volume.
- To improve the monitoring of distibuted services in a heterogenous environment.
- This will serve as a foundation to build an eco system where researchers can share data and new algorithms/tools to analyse data efficiently and help each other to advance research further.

3 Scope

Describe the areas expected to be covered or impacted by the proposed activity, such as organisational areas, systems, processes, resources.. i.e. what *is ‘in scope’. This is not a list of what will be done but identifying the services, areas or what, will be affected.*

Also please enumerate specific items which although they could perhaps be related are intentionally not addressed by your proposal ("Out of Scope").

1. In Scope

- The project will build up a distributed storage as well as computing system to analyse data at large scale. The architecture will be horizontally scalable and based on commodity servers.
- The project will demonstrate a demo service using the build system to analyse netflow/genomic data.

2. Out of Scope

- The project will not try to invent its own specific machine learning algorithm. Although it will provide a framework where researchers can build their own algorithms to analyse data.

4 General Information

Outline any potential issues, risks, dependencies, assumptions, constraints and limitations or any other points that may be useful to help assess the proposal.

1. The project will depend upon resources from an IaaS platform e.g. Uninett internal IaaS or Nordunet IaaS.
2. The project in first phase will not be a production service but a prototype to show the possibilities in this area which may open future possibilities for colloboration to provide such a service at an inter-NREN scale.