**EGI Solution:**

**High-throughput Data Analysis**

**Executive Summary**

The “High-throughput Data Analysis” solution is aimed to help individual researchers, and research communities regardless their size that need intensively using data management and computational capacity in a distributed infrastructure. The challenges that they typically face are limitations in the capacity, and technical and administrative problems for sharing their data sets, or accessing different capabilities in different locations. This may have cost

With this solution they gain access to distributed and virtually unlimited resources, which allocated via a central process, and with uniform interfaces. They can also be authenticated by a single process in different remotely located sites. They are enabled to perform their investigations in a collaborative way, which otherwise would not have been possible. And they can do it effectively and cost-efficiently.

The solution is built with a combination of services already provided by the EGI.eu organisation, such as Operations, Technology and Security Coordination, and Technical Consultancy and Support, but also using the resources located in the federation that comprises the European Grid Infrastructure.

This solution is one of the ways in which EGI attends to the needs of the researches and research communities not only within the European Research area but also located in other main research areas.

**Table of Contents**

1 Target Groups and Specific Challenges 4

1.1 Target Groups 4

1.1.1 Research Infrastructures and Resource Centres part of EGI 4

1.1.2 Resources Centres aiming to integrate with EGI 5

1.1.3 Resource Centres aiming to form an independent federation 5

1.2 Specific Challenges 6

2 EGI Solution 7

2.1 Objectives 7

2.2 Accessing the Solution 7

2.3 Building the Solution 8

3 Value Proposition 10

4 Key Performance Indicators 11

5 SuCCESS Stories 11

6 Conclusion 11

# Target Groups and Specific Challenges

## Target Groups

The “High-throughput Data Analysis” solution targets groups of researches who work on the same topic, or share data or application software. The researchers need to analyse or produce large set of data through the execution of large ensembles of loosely coupled computational tasks, which can be either single processes or parallel processes. The user community need to access the distributed resources and datasets in a collaborative way.

## Specific Challenges

The challenges can be described as follows:

* Users do not have access to enough resources within their institution.
* The user community has resources distributed in different resource centres, and they need to have uniform access to them.
* The access to the distributed resources and datasets must facilitate the collaboration among the researchers.

If these challenges are not properly met, the research-users cannot access the capacity they need to produce the results in the expected timeline. If the computing or storage capacity is available to the users but not uniformly accessible because it is served by different providers, the users may be forced to use additional resources for accessing them, and adapt the data analysis workflow depending on the site being used at the moment.

# EGI Solution

## Objectives

EGI offers the “High-throughput Data Analysis” solution to enable users to:

* Access transparently distributed resources with uniform interfaces,
* Be authenticated in an uniform way in different sites,
* Organise user communities to regulate access to services and data throughout the infrastructure,
* Access resources assigned through a central allocation process.

Users are able to access distributed resources through common standard interfaces uniformly available in the different resource centres. Independently from the Local Resource Management Services chosen by the administrators of the different sites, the user can manage his data and execute and control the computational tasks using common services and APIs.

Users’ identity is uniformly recognised in the whole infrastructure, making possible –as an example– for a computational task running in one resource centre to access data stored by the user in another.

EGI provide services to organize a research organisation, grouping users and assigning them specific roles. These additional attributes can be used in the authorization process, to regulate the capability of the users to perform specific actions or to access a dataset, based on their role in the organisation.

EGI provide a single entry point to a federated pool of resources that can be allocated to new or existing user communities who need resources to perform their research activity.

## Accessing the Solution

Users can access resources through common interfaces and common client commands.

Users usually access the computing and storage services from a user interface (UI), which is a machine (virtual or physical) where they can have shell access. UI installation contains the client for most of the services deployed in EGI, the clients allow users to submit computing tasks, retrieve results, store and manage data. EGI provides virtual images and installation tools to deploy locally user interfaces, even on a laptop machine.

To authenticate in a uniform way on all the EGI services users need to have grid credentials, which are x509 certificates[[1]](#footnote-1). Users usually keep grid certificates password protected on the UI they normally use.

Users without a certificate can ask one to their Certification Authority (CA), or access the EGI catch-all CA through the Registration Authority distributed in different countries. More information is available on the EGI website[[2]](#footnote-2).

Authorization on resources is normally regulated by Virtual Organization (VO) membership. Users who are not members of any VO can join an existing VO that supports the research topics of the user, or create a new VO and have it approved as production by EGI. Clearly individual users, or small users groups, may prefer to join existing VOs to reduce the overhead of creating a new VO.

Users who have federated their resources in EGI can, with the credentials and the VO, start accessing their distributed resources.

Users who do not own resources can contact the NGI or NGIs individually to ask for opportunistic usage of the available resources, in practice this will result in enabling the VO in the services operated by the NGI. A second option is to ask for resources directly to EGI, in order to have access to the pool of resources devoted to central allocation, these resources are allocated centrally with an open call that can be accessed by all users.

## Building the Solution

The “High-throughput Data Analysis” solution builds upon the middleware services that enable access to distributed resources.

Services can be grouped in three categories based on the services and resources accessible: compute resources, storage and data management resources and cloud resources.

For user authentication EGI services enable grid certificates certified by the Certification Authority (CA) members of the EUGridPMA organisation. EGI in collaboration with EUGridPMA deploys and tests the CA certificates distribution in all the production services, to ensure uniform user authentication.

EGI in collaboration with the NGIs deploys a number of VOMS services, which are used to store VOs information, including users’ membership, grouping and roles within the VO. Users organised in a VO can request access to one VOMS service and register – and manage – their members.

All the services described in this section are provided by several Technology Providers (TP). They release independently their software in different distribution channels. EGI collects the most relevant components and release them in dedicated repositories as part of the UMD[[3]](#footnote-3) distribution. UMD releases are tested at several levels including a pre-production deployment. Sites and users have the added value of having all the components that build the infrastructure accessible in a single repository and with an additional quality assurance process on top of what is performed by the TP.

Operations coordination

Technology coordination

Security coordination

Federated operations

Helpdesk support

Technical consultancy and support

Repository of validated software

# Value Proposition

Easy access to shared computing and data services from independent resource providers where to provision owned resources or access unused ones in a uniform way and preventing single vendor lock-in while optimising utilisation.

| **Problem** | **Provided solution** | **Added Value** |
| --- | --- | --- |
| Users do not have access to enough resources within their institution. | Common Core Infrastructure Platform based on standards, common interfaces and protocols, communication, planning and coordination | * Transparent access to distributed computational infrastructure beyond the local capacity restrains
* Flexible use of data storage and computation across disciplines and borders
* Easier management of access to services and data throughout the whole infrastructure
* Access to resources assigned through a central allocation process
* More efficient use of available resources, both computational and human
* Time and effort saving, more efficient research process
* Improved user experience
 |
| The user community has resources distributed in different resource centres, and they need to have uniform access to them. | Centrally-provided expertise and streamlined best practices on how to set up and manage federations |
| The access to the distributed resources and datasets must facilitate the collaboration among the researchers. | Federated service management best practices, cost-effective sharing of services (support, processes, policies, activities), community expertise & re-use of tools/output from public funded projects |

# Key Performance Indicators

The following Key Performance Indicators (KPIs) are identified to measure the success of the solution.

* Number of researchers / research communities using the system,
* Number of new users in a given period of time,
* Number of jobs run/ core time / HEPSPEC06 hours.

# SuCCESS Stories

High-throughput data analysis services have already being used by the scientific community for a few years and there are already some good cases that can be presented as success stories.

The application of the services provided by EGI and facilitated by this solution allowed applying a 3-D model system to study the effects of a toxic spill in a real tailing dam. The research involved complex calculations based on 15,000 different sets of input parameter values, which was at their time considered very demanding in terms of computation resources. The results were considered extremely satisfactory by the researcher and other stakeholders.

A very singular case comes from the world of mathematics. A very computational demanding study was designed to deep on the understanding of the famous Goldbach’s conjecture, which still remains as one of the biggest challenge among mathematicians. This project submitted 173,816 jobs to the European Grid Infrastructure adding up to 869,080 tasks, consuming 1 CPU-hour each, which took 7 months to process. The same task in a normal laptop would have needed about 99 years.

# Conclusion

The “High-throughput Data Analysis” solution is a key part of EGI’s solution portfolio. It is aimed specifically at helping researchers with most of the problems in using data management and computing resources that the EGI has identified as being more problematic.

With this solution researchers and research communities gain seamless, transparent access to larger computational capacity, in a way which allows them to concentrate on their own business, obtaining research results in a secure and more swift way. They are also enabled to work collaboratively with other groups located remotely, across countries, and even continents.

1. <http://en.wikipedia.org/wiki/X509> [↑](#footnote-ref-1)
2. <http://www.egi.eu/how-to/get_a_certificate.html> [↑](#footnote-ref-2)
3. http://repository.egi.eu/ [↑](#footnote-ref-3)