PROJECT FINAL REPORT

Grant Agreement number: 619439

Project acronym: VELASSCO

Project title: Visual Analysis for Extremely Large-Scale Scientific Computing

Funding Scheme:

Period covered: **from 01/01/2014 to 31/12/2016**

Project co-ordinator name, title and organisation: Dr. Abel Coll, Principal Investigator, CIMNE

Tel: +34 93 401 74 03

Fax: +34 93 401 74 41

E-mail: projectes@cimne.upc.edu

Project website address: http://www.velassco.eu/

Table of contents

1 Final publishable summary report

1.1 Executive summary

As the amount of data obtained with state-of-the-art engineering application simulation tools is growing fast, Big Data and high-performance computing (HPC) are converging to face the same technological challenges. It is clear that by 2020, it will not be possible to store the most crucial simulation results, such as the ones coming from the aeronautic and automotive industry or from the prediction of natural disasters, in a single machine or server.

One of the problems that engineers and scientists running high-end real-scale simulation models have to face is directly related to the management and exploitation (in terms of visualization) of the outputs of their models. In fact, how to scrutinize the data that are distributed over heterogeneous clusters of servers, located remotely, has become one of the key difficulties to be overcome to ensure that users (both in academia and industry) are able to take the best advantage of current (and future) HPC infrastructures.

VELaSSCo has developed a new concept of integrated end-user visual analysis methods with advanced management and post-processing algorithms for engineering modelling applications, prepared to scale up to real-time petabyte level of simulation data. The VELaSSCo Platform enables a real-time interrogation of simulation data and advanced analysis and visualization of these simulation results in a smart and user-friendly way, providing key engineering information.

Main concerns have to do with the handling of large amounts of data of a very specific kind, intrinsically linked to geometrical properties:

- how to store, access, simplify and manipulate billion of records (ranging from tens or hundreds of terabytes to even petabytes);
- how to extract only the relevant information and represent it in a feasible and flexible way; and
- how to visualise and interactively inspect the huge quantity of information they produce taking into account end-users' needs.

VELaSSCo achieved this by a synergistic collaboration between experts with relevant and diverse background including Big Data handling, advanced visualization, engineering simulations and standardization; together with a significant user panel with a potential membership of over 100 scientists and industrial R&D engineers, which includes members from research organisations, SMEs and companies spanning key European industrial sectors including aerospace, household products, chemical, pharmaceutical, forestry and construction.

As the main result of the R+D work developed in this project, the VELaSSCo Platform has been released. It has been tested and evaluated by our user panel, resulting in some promising expectations. From the union of the most extended Big Data technology based in the Hadoop ecosystem and the most consolidated engineering simulation techniques based in the finite and discrete element method, the project has made possible to shorten the gap between them, and to increase the know-how on the key characteristics of Big Data technologies to be used for simulation data.

The VELaSSCo Platform, with its two options of Open-source and Closed-source implementation, has been made public for the scientific community through a GitHub repository (https://github.com/velassco/VELASSCO). The main modules of the platform are open source, and its architecture is intrinsically modular. So the end of the project, rather than represent the end of a development, opens the door to new implementations and developments based on the designed Big Data architecture for engineering simulations.

1.2 Summary description of project context and objectives.

*The **Vision** of VeLASSCo is to provide new visual analysis methods for large-scale simulations serving the petabyte era and preparing the exabyte era. It does this by adopting Big Data tools and architectures for the engineering and scientific community and by leveraging new ways of in-situ processing for data analytics and hardware accelerated interactive visualization.*

Today, numerical simulations are widely used in scientific and engineering fields to lead to a better understanding of physical phenomena and create better design solutions for given problems. Commonly, the results of a simulation come attached to a mesh (used for the calculation). To analyze them, engineers and scientists use post-processing tools because results may come in a raw format, and sometimes some results derived from them are also of interest.

These tools extract information from the meshes using a variety of techniques that allow a better understanding of the results. This process not only consists of accessing the results provided by the simulation, but also of generating new results from it, by applying **post-processing operations**. Examples of such operations include calculating the 0-level iso-surface of the air pressure in an hydrodynamic simulation, cutting through the inner part of the domain in several planes, or calculating streamlines of the velocity. We classify these post-processing operations to be part of the *data analytics* domain (used in other fields).

All this information has to be visualized in a user-friendly way by rendering and presenting it to the user in an interactive 3D environment. This part is the so called **visualization** of the results. A post-processing system is covering these two parts: post-processing operations and visualization.

As an example, the image sequence shown in Figure 1 corresponds to the simulation of sea waves against an oil platform structure. The raw results coming from the simulation are the velocities and pressure values at each node of the tetrahedral (volume) mesh representing the domain.

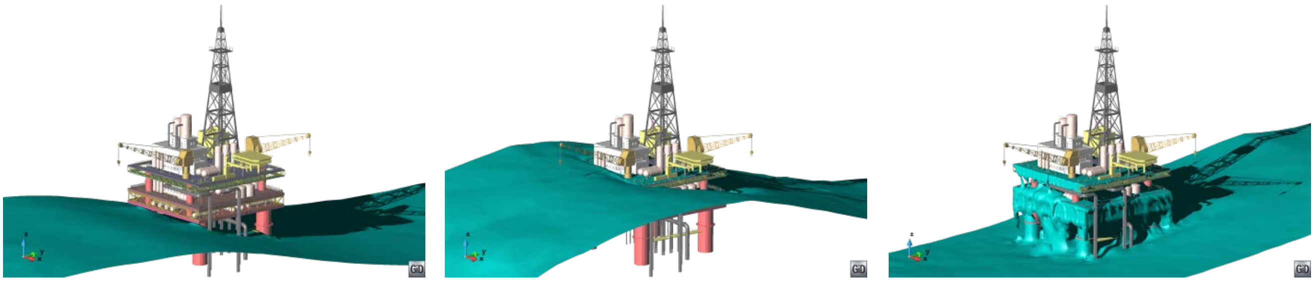

Figure 1*: different stages of an animation of the iso-surface representing the sea surface.*

The free surface of the sea is represented with a surface mesh, which is not provided directly by the simulation: it is obtained in a post-processing operation from the tetrahedral mesh that defines the domain, by calculating the iso-surface at pressure equal to 0 (the physical condition a free water surface has to meet). This is the post-processing operation. It can also be seen that a rendering process has been applied to this surface mesh in order to see it smoothly, and to project the shadow of the platform onto it: this is the visualization part. This combination of post-processing and visualization has to be conducted for each time step of the simulation.

It is important to note that some results coming from the post-processing operations (like iso-surfaces or cut plane meshes) have a very short life span and are only visualized when the user requests this information until the next request or until the end of the visualization session.

As computers increase their memory and number of cores (CPUs), with the High Performance Computing (HPC) resources at the top of the computation capabilities, the size of the simulations grows exponentially in order to solve more complex problems and deal with large-scale challenges such as predicting the pollution of a large city under certain climate conditions.

This requires extremely large amounts of results that need to be post-processed and visualized. Simulations at this scale commonly exceed the storage capacity of a single computer, so the geometry of the domain must be split into different parts and distributed over several machines (see for instance the example of a big telescope in Figure 2).

Most simulation solvers are already prepared to run in such a Distributed Memory Parallelism, but this is typically not the case for post-processing systems.

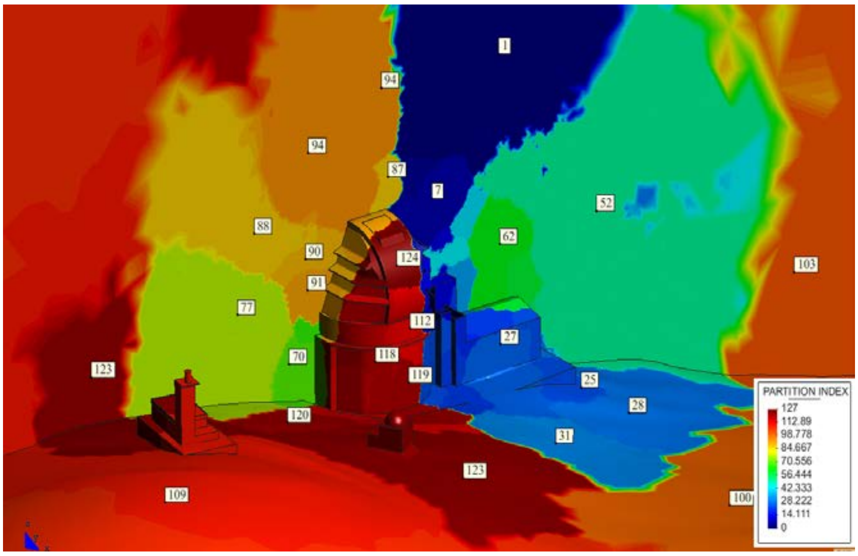

Figure 2: *Example of the model used for a simulation of the air surrounding a big telescope. View of a cut of the model. The different colours are the sub-domains in which the model is subdivided to run the simulation.*

There exists a similar, well known problem with popular social networks, where millions of users continuously generate large amounts of new content that has to be distributed over thousands of servers to be available 24x7.

The main **objective of VELaSSCo** project is to build the VELaSSCo Platform, a system that performs distributed **post-processing operations and visualization of very large simulations**. To reach this goal we will tap into recent advances of Big Data field, which is still unexplored in simulation.

To address this objective, VELaSSCo brings together **Simulation** and **Big Data** in a robust and fast platform that allows us to access extremely large amounts of distributed data from very complex virtual models with structured and closely linked data and from measured heterogeneous and loosely coupled data.

It's important to remark that the advances in the use of well-established and common Big Data technologies (basically based on Hadoop) for the VELaSSCo platform will simplify the adoption by the simulation engineers of these Big Data technologies, and extends the Big Data market by including the CAE domain, what can be considered an important output of the project.

The design of the VELaSSCo Platform has to account for several engineering challenges:

- User requirements and assessment: demand for real-time visualization, fast interactivity and interoperability, realistic visualizations, and so on. (WP1, WP5)
- Simulation data characteristics: very big set of structured and tightly coupled data, highly distributed, possibility to access the data during ongoing simulation, and so on. (WP1)

To address these two challenges, VELaSSCo uses the following set of strategies:

- Take advantage of Big Data technology (in a broad sense, architecture, methods, philosophy…) to perform the **post-processing operations faster and in a distributed way**. As Big Data works with highly scalable parallel systems, the post-processing algorithms should be programmed to fit these characteristics (WP2, WP3).
- Take advantage of the Big Data properties to **pre-compute and store** some of the more expensive post-processing operations. Due to the amount of storage facilities the Big Data can manage efficiently, some of the common post-processing operations can be precomputed before the user demands them (WP3).
- **Reduce and simplify the data to be visualized**, in order to reduce the time for sending the information to the client. Although the results of a simulation may be onto a volume mesh, the user may want to visualize datasets in surface or line meshes (like iso-surfaces or streamlines). This aspect already reduces drastically the information to be sent to the visualization client. Furthermore, novel mesh simplification algorithms are studied in the project to simplify these meshes. Also, the use of LR-BSplines is studied to represent in a more compact way the sets of results, reducing even more the data size (WP3, WP4).
- Use new **GPU-based algorithms** and formats to achieve interactive visualization rates in the visualization client.

To illustrate these aspects, a zoom of a racing car model with a contour fill of pressure is shown in Figure 3. In this example, a Computational Fluid Dynamics (CFD) simulation is done of the air surrounding the car. The calculation mesh has 104 million tetrahedral and the skin of the car has 6 million triangles. This simulation used 96 computing cluster nodes in an HPC infrastructure. The mesh was split in 96 different parts and no single computer had the whole model stored.

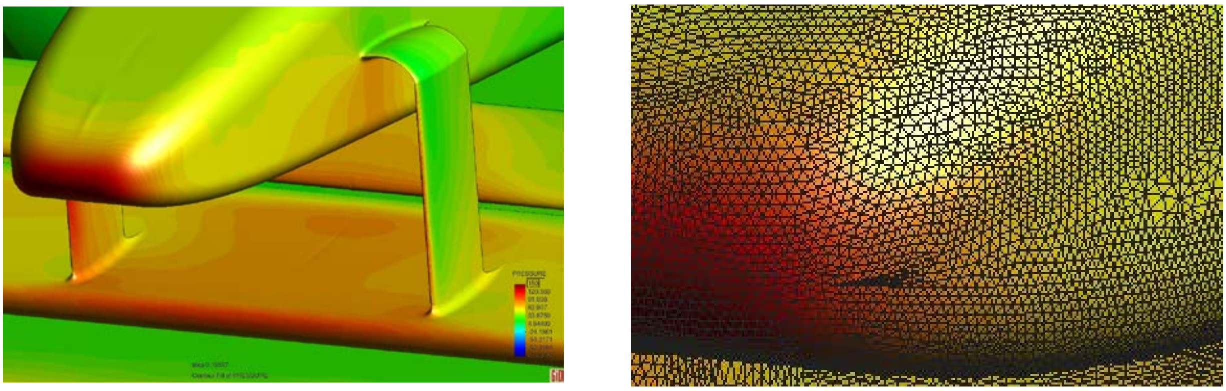

Figure 3 *Contour fill of pressure on a part of a racing car and a view of the calculation mesh.*

As a first point, a user might be interested in knowing the pressure in the skin of the car (not in the whole domain). Only by considering this, we avoid to transfer data of 104 million tetrahedral elements to the visualization client.
Second, the level of detail of the calculation mesh can be appreciated in the second image. A mesh to represent the colour scale for visualization purposes can be coarser, so by applying a simplification algorithm to the mesh, we can reduce the data to be transferred even more.
Once the (already reduced) data is in the visualization client, taking advantage of the novel GPU-based techniques, the user will be able to visualize the model with the contour fill of pressure in an interactive way on a single computer, although the results coming from the simulation needed an HPC centre to be generated and stored.

1.3 Main S&T results/foregrounds

The main result of VELaSSCo Project is the first functional prototype of the VELaSSCo Platform: a platform, based on big-data technologies, to post-process and visualize large distributed data sets coming from numerical simulations.

To reach this goal, the consortium has relied on the following main pillars:

- A **User panel** has been created in order to take into account the final user, not only in the design phase of the Platform (obtaining the main requirements for the Platform), but also to feed the iterative process of development with its feed-back (including evaluation methodologies in the intermediate stages of the implementation). It also has been crucial to consider the user panel for defining an exploitation strategy of the product as much close to the market as possible. (WP1)
- A **big data architecture** with two options, one for Open-source and one for Closed-source implementation, have been designed considering the main user and technical requirements. This has been one of the key points of the project, as this architecture is the core of the platform, and is the milestone representing one of the main aspects of the project: the union between big data and numerical simulations worlds. The architecture of VELaSSCo Platform uses big data tools and paradigms, but adapted to the numerical simulations needs. (WP2)
- An exhaustive **implementation plan** has been designed and followed by the consortium in order to develop and implement the set of functionality (VQueries) for the VELaSSCo Platform. It has to be highlighted that the development of the platform has been done completely collaboratively, integrating developments of every partner of the consortium. (WP3 and WP4).
- An **evaluation methodology** for the platform has been designed based on the Goal-Question-Metrics (GQM) and SRVQual methodologies, considering the following main dimensions: architecture, algorithms, effectiveness and usability. It has been applied in two evaluation events (one for the first prototype and other for the final version of it), where members of the user panel has used the VELaSSCo Platform following some predefined use cases. The feedback from these evaluation events has been crucial in order to prioritize the implementation plan and include slight modifications in the big data architecture defined in the beginning stage of the project, always having in mind the end user vision.

In the following subsections, each one of the mentioned pillars is detailed.

1.3.1 Kind of data used

The aim of this subsection is to put in context about the kind of data used in VELaSSCo Project.

Among all the kind of numerical simulations, in the frame of the project we have focused on the ones using Finite Element Method (**FEM**) and Discrete Element Method (**DEM**).

Finite Element Method is a computational approach, which uses solvers to compute solutions to partial differential equations on a domain defined by a mesh, which is a discretization of the domain. For example for a car simulation, the domain can be the air surrounding the car. The result of these kinds of simulations is a mesh with attached results in each time step solved by the simulation engine. It has to be considered that the mesh can be different in each time step, but the use cases selected in VELaSSCo platform are using the same mesh for every time step.

In DEM simulations, the domain where the simulation is applied is discretized by particles. The particles interact with their neighbors through contact forces. The data of DEM simulations is very dynamic because particles and contacts can be created or lost during the simulation. Although DEM simulations can use different types of particles (different geometries of them), in the frame of VELaSSCo project only spherical particles have been considered.

1.3.2 User panel

A User Panel including scientists and industrial R&D engineers from universities, research centres, SMEs and companies from key European industrial sectors such as aerospace, household products, chemical, pharmaceutical and civil engineering was created at the start of the project and has seen continuous growth in numbers..

Since the beginning of the project, a special attention has been given to the user panel, in order to:

- Define the end-**user requirements** (visualization and data analytics) of the system to be developed. The user requirements have formed the basis for drafting the technical requirements to be implemented by the VELaSSCo Platform.
- Provide the scenarios where these requirements should apply, and select the examples to be used along the project lifetime. Some of the members of the user panel have played the role of **data providers**, allowing the consortium to have real complex simulation models to test the Platform in all its dimensions.
- **Test and evaluate** the VELaSSCo platform in the first and final stages. The evaluation of the platform has been done according to a predefined methodology (see Section 0) in the frame of evaluation events, where the members of the user panel use the VELaSSCo Platform, and provide the consortium with the feedback about it.
- Guide the **exploitation plan** (both of the consortium and the individual partners) considering the current needs of the market.

It has to be considered that, during the lifetime of the project, some of the aspects mentioned above have suffered changes (due to new technologies adaptation, changes on priorities of institutions, etc…). This has implied an iterative process where the consortium has had to adapt some of the decisions taken accordingly to the new scenarios presented. This is a natural process considering software development when cutting-edge technologies are used.

The user panel has been organized in four main engagements:

1. The first engagement for defining the initial requirements for the users (details in deliverable D1.1).
2. The second engagement for analysing the status of the development of the system and the architecture, and define potential modifications and corrections in the system (details in deliverable D1.2).
3. Testing of the initial prototype of the system and final corrections/modifications (Deliverable D5.8).
4. Presentation and testing of the final prototype (Deliverable D5.9)

Beside these four main items, the user panel has been updated regularly about the status of the development and implementation of the platform.

Apart from the regular communications, a specific meeting was held with the associated T-MAPPP (a European research project) partners. The meeting took place on 2014, March 18th in Edinburgh, and the main goal of the session was to obtain the DEM/ CFD-DEM user requirements for VELaSSCo platform. A total of 16 DEM/CFD-DEM users from different Universities, research centers and companies attended to the event.

The number of members of the user panel has grown more or less homogeneously from 24 (February 2014) to **59 members** at the end of the project (December 2016). Currently, there are members from **35 different organizations**, with two thirds belonging to academic institutions, and one third to industry. The distribution of members according to the nature of their institution is depicted in Figure 4.

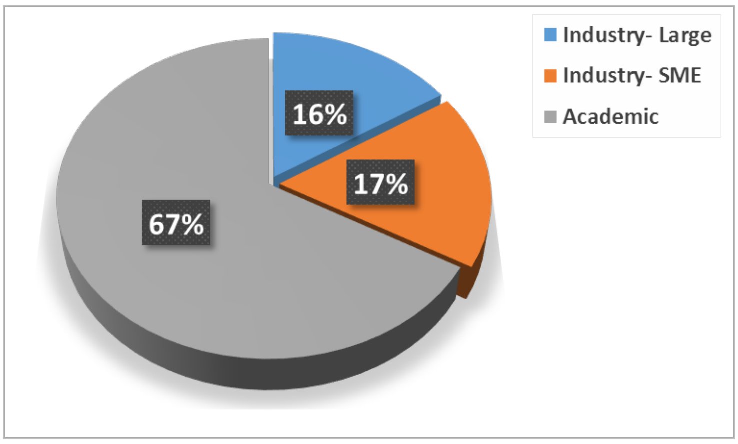

Figure 4. *User panel distribution according to the nature of the institution.*

User panel members are mainly working on processes or scientific research involving numerical simulations, and they can be distinguished into two basic profiles:

- Institutions requiring a post-processing and visualization platform for very large simulations
- Institutions dealing with not so large simulations, but foreseen to increase the size of data to be processed and visualized in the near future.

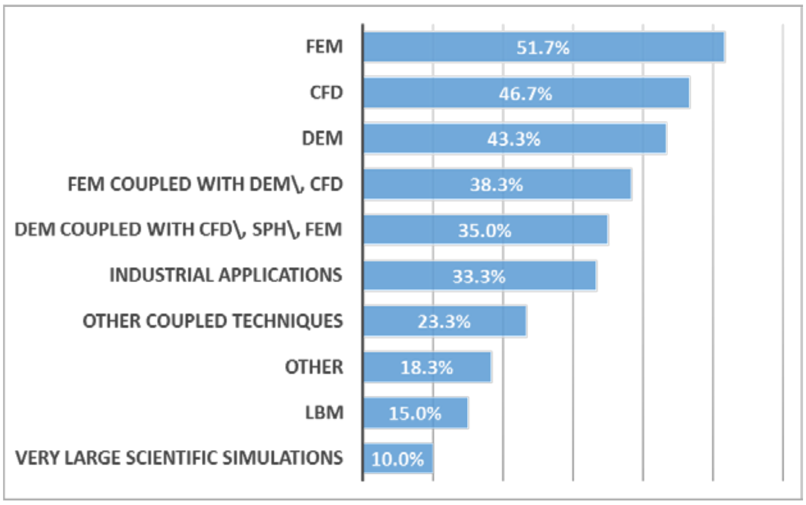

Figure 5. *Kind of simulations used by the user panel members*

According to the kind of numerical simulations, the distribution of the user panel members can be depicted in Figure 5. It has to be noted that the fields of this graph are not complementary (CFD simulations, for example, can be done using FEM or DEM methods), nevertheless, this information provide us with a general view of the wide range of applications of such a platform as VELaSSCo.

1.3.3 Big data architecture

As mentioned above, one of the key pieces to reach a successful release of the VELaSSCo Platform is to design an appropriate software architecture that satisfies the requirements (user and technical). The design of this architecture has been the main objective of the WP2 of the project, named *Data formatting, handling and storage*.

The first stage of the project has been devoted to perform a very deep and extensive study of the state-of-the-art in big data tools and technologies, present in the deliverable D2.1.

The work done in the frame of VELaSSCo has been pioneer at the time of merging big data and numerical simulation fields, and a lot of effort has been devoted to try to fit big data paradigms into simulations needs. Within the consortium, there are partners with high expertise in both worlds, and it has been really exciting to know each other, and further explore some aspects related to the new point of view tackled by the project.

The main aspects that were considered when analyzing the current big data solutions are:

- Highly distributed systems (including HPC, Grid and Cloud computing):
- Current characteristics of Cloud infrastructures in terms of storage, computation and network intercommunication;
- Hardware requirements
- Hadoop ecosystem. As Hadoop can be considered the kernel of big data, an important focus has been devoted to understand it, and to know how further we can take it considering simulations requirements.
- Applicability to commercial and standards based database solution. For the Closed-source option of the architecture, focus has been on how to integrate a commercial tool such as Express Data Manager (EDM) from Jotne, to handle the VELaSSCo Big Data requirements.

Nowadays, almost all the simulation codes dealing with very large simulations are using HPC facilities. Although both HPC and Cloud infrastructures work in a distributed way, HPC is more focused in intensive distributed calculations (CPU requirement), and leverages the high speed network connection between compute nodes.

Discussions with the partners highlighted some key requirements: an efficient computational model, a distributed computational model, a simple interface to access data, a computational model suitable for batch and online queries and local distributed storage to speed-up data accesses. Considering these aspects, it has been decided to design the prototype of VELaSSCo Platform for running on a HPC infrastructure.

Hardware requirements

Specifically, the EDDIE cluster has been used in the frame of the project. A schema of the resources of the EDDIE cluster devoted to VELaSSCo project is depicted in Figure 6.
The hardware configuration per node of the EDDIE cluster is detailed hereafter:

- CPU: 2 x Intel Xeon E5-2630v3 @ 2.4 GHz (16 cores)
- RAM: 64 GB RDIMM, 1600MHz RAM
- Hard disk: 1 x 1.2TB

At the end of the project the consortium could take advantage of 18 extra nodes with the additional local storage available on each node, summing up to a total of 38 nodes and ~110 TB.

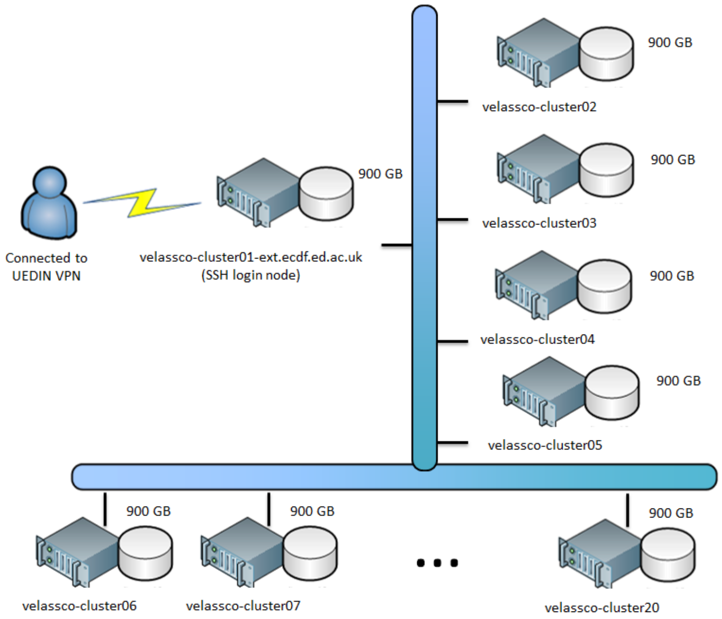

Figure 6. *EDDIE HPC cluster resources dedicated to VELaSSCo Project*

In the early stages of the project, as the EDDIE cluster was not available yet for the consortium, CIMNE gave access to all the partners to its internal cluster (Acuario). It is an older and less powerful cluster, but it has played a key role for the first developments and implementations of the platform, as it has permitted to begin the implementation tasks at time. A schema of the resources of the Acuario cluster devoted to VELaSSCo project is depicted in Figure 7.

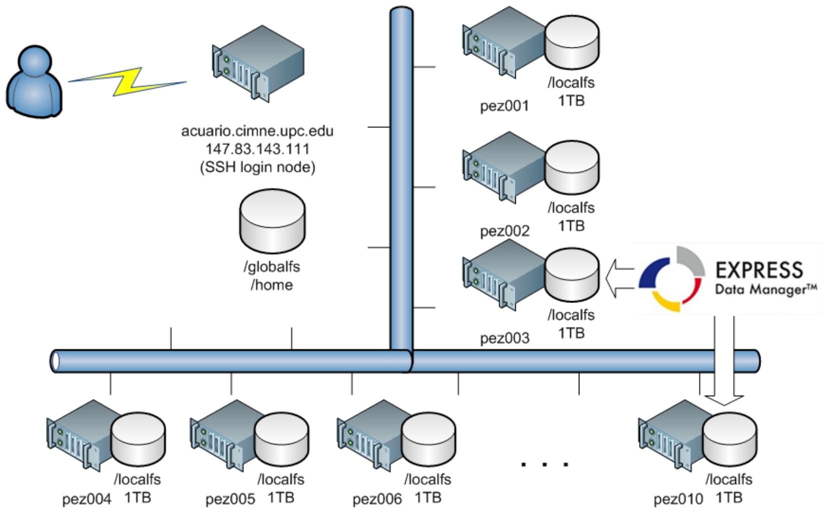

Figure 7. *Acuario cluster resources dedicated to VELaSSCo Project*

Hardware configuration of the Acuario cluster is detailed hereafter:

- CPU: 2 x Intel E5410 @ 2.33 GHz (total of 8 cores)
- RAM: pez001: 32GB, the rest: 16 GB
- Network: DDR x4 Infiniband (20Gb/s) + 1Gbps Ethernet
- Hard disk: 1 x 150 GB + 1 x 1TB

A common mistake with Big Data is to try to provide a solution that services simultaneously HPC and Big Data (computation and storage are located on the same cluster) applications. Requirements of both systems are not similar. HPC applications focus on high intensive computation and network interaction but centralized storage and BigData focus on not so intensive computation and network interaction but on high distributed Input/Ouput bandwidth, by means of local storage and data replication. An optimal solution would have been to go for a dedicated Big Data post-processing platform. But to be suitable for any existing architecture and deployable on any kind of IT system this architecture should be flexible and scalable, eventually using commodity nodes with virtualization. With this method, VELaSSCo targets the largest architecture at the lower price.

The VELaSSCo platform can be deployed in a network of commodity nodes with local storage, even on HPC clusters. The VELaSSCo platform, both its Open-source option and the Closed-source option, have been successfully deployed on CIMNE's in-house cluster Acuario and on UEDIN's cluster EDDIE.

Platform architecture/software design

Building this platform can be accomplished with two strategies: use an existing framework or develop a new framework from scratch. The first solution seems to be the more suitable one; development cost and complexity being drastically reduced. Moreover, with an existing framework, it is possible to use existing plug-ins. The chosen architecture must be modular, following the plug-in concepts, and has to fit the proposed requirements: to be deployable on an important set of nodes, and support for virtual machines. The solution has also to be suitable with HPC facilities. For an existing solution, it is necessary to choose an evolving framework with regular updates. Using a deprecated solution can bring many design and implementation issues, increase the amount of work and result in poor performance.

The Cost of the platform must also be considered. Linux seems to be the most suitable operating system, where lower execution times are achieved. With this strategy, the global cost of our platform will be drastically reduced. It can be deployed on a larger cluster (cost will be mainly dedicated to hardware).

Two big data frameworks are on the table: one brought by one of the partners and another open-source. Both should share the skeleton of the architecture, whose specification is presented in deliverable D2.2. The main idea is to take advantage of both, of commercial modules, like EDM object oriented database, and of standard open source software.

From these elements, the most suitable solution for the VELaSSCo Platform has been identified as the **Hadoop ecosystem** (details in D2.1) with some extensions.

EDM database

Express Data Manager™ (EDM) (see http://www.jotneit.no/products/express-data-manager-edm for more information) from Jotne EPM Technology is an object-oriented database that is well suited to handle both structured and unstructured FEM and DEM data. The reason for this is that ISO has standardized an object oriented schema for storing and exchange of FEM and CFD data, ISO 10303-209, Multidisciplinary analysis and design (AP209).

At the consortium level, it has been agreed to adapt the architecture of the VELaSSCo Platform to integrate the EDM database for the storage of the data, at first. This integration has evolved during the project including some VELaSSCo functionality and analytics as EDM port to Linux matured. This ended up in two options of the architecture of the VELaSSCo Platform: one using OpenSource tools from the Hadoop ecosystem, and the other one integrating the EDM database (a proprietary software). Both options of the platform are usually referred as the 'Open-source' and the 'Closed-source' options of the VELaSSCo Platform.

VELaSSCo Platform architecture

A detailed explanation of the architecture chosen for VELaSSCo Platform is provided in deliverable D2.2 – *Specification of the Big Data Architecture*.

The architecture of VELaSSCo Platform has been based on Lambda Architecture (LA) proposed by Marz [2]. A schema of it from a service point of view is presented in Figure 8.

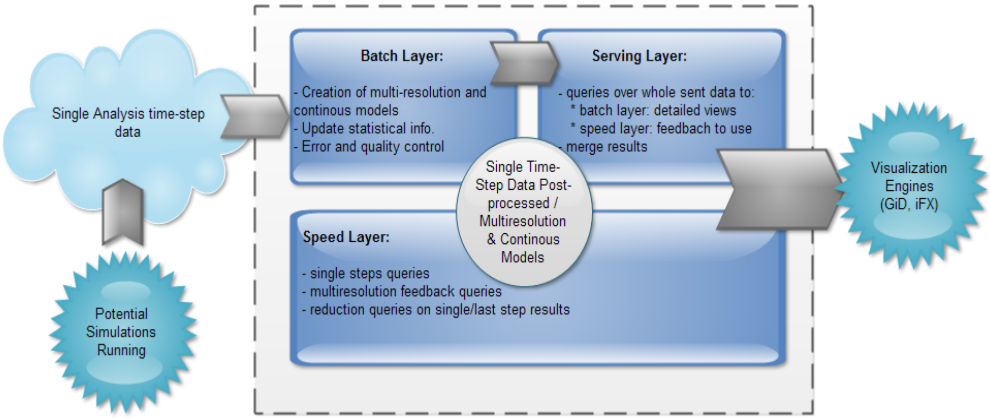

Figure 8: *VELaSSCo Lambda architecture schema*

It primarily consists of three layers: batch, speed, and service layers. The roles and functions of these layers are summarized as follows:

1. The engineering data from the simulation are constantly sent to the HPC clusters with the help of the petabyte sized data interface, thus creating a constant growing master dataset. With this constantly growing dataset, the batch layer starts.
2. The batch layer pre-computes query functions (based on the domain knowledge) on subsets of data of the master dataset, and then stores the partial results for a later dispatch, thus lowering the system response time.
3. As soon as the request for the visualization enters the service layer, the database fetches appropriate computed partial results. These results are sent to the clients that render the final images, using the locally available GPUs.
4. While the batch layer performs its pre-computations, if any new data arrives they are sent to the speed layer. The speed layer runs incremental algorithms and helps in visualizing only the real-time (recent) data.
5. Visualization queries are resolved by getting results from both the speed and batch layers.

The architecture defined has considered the following requirements from the scientific visualization and big data field:

- Be effective on most scientific hardware.
- Deal with data location and storage considering the simulation engines characteristics.

- Query engine layer able to work at two levels: fast and heavy queries treated by different layers, ensuring an interactive use of the platform.
- Support extensibility considering the addition of new modules/tools in the future.

This conceptual view has been summed up in the platform architecture depicted in Figure 9, where two layers encapsulate different functionalities: data management, process and analysis, and external interactions:

- Data Layer: responsible to store, access and translate the data and will answer the data queries coming from the "Query Engine Layer". Eventually, when data is injected in the platform, it may trigger simplification queries, to create coarser views of the simulation data, and other queries;
- Engine Layer: receives user requests (VELaSSCo queries) from the visualization client, process them, generating data queries to the Data Layer processing and formatting these results in a GPU-friendly way for the visualization client.

This approach allows the transparent exchange of the Data Layer between the Hadoop and EDM version, for the Engine Layer. In addition to these two layers, two libraries/modules enable the connection between the platform and the final user:

- Data Injection library: provide a mechanism to inject data from the computer engineering simulation solvers to the VELaSSCo platform. After establishing the communication with the Data Layer, including user validation, the injection module sends the data to the Hadoop Abstract File System. The Data Layer stores the data in P4 / GiD format if the HDFS is used as storage or in an ISO STEP compliant way in the EDM-DB system with the option to translate the data to STEP/AP209 when using a corresponding EDM-plugin.
- Platform Access library: communicate with the visualization client and translate user requests into VELaSSCo queries, receive the results and show them to the user. The module also generates VELaSSCo queries automatically, for instance when the user just navigates through the simulation model, or to ensure interactivity it will handle coarser models and issue full-detailed queries when the user zoom on parts of the simulation model.

Inside the Query Engine layer the following modules have been developed:

- *Query Manager*: manages the VELaSSCo queries, evaluate their execution costs and provides feed-back to the visualization client for heavy queries. The Query Manager also synchronizes and controls the visualization parameters and issue visualization queries to the Real Time module, like the 'traveling' through the model for an interactive navigation.
- *Analytics module*: performs data analytics over the stored simulation results, like calculating the skin of a volume mesh, stream-lines, iso-surfaces, etc. This module also generates coarser versions of the simulation data in order to provide fast feedback to the user and the Discrete to continuum transformation needed by the DEM simulation data.
- *Batch module*: performs heavy data analytics and provides feedback on the progress of the query.

- *Real Time module*: provides feed-back to the user whether in full detail (for instance of detailed views of navigation through the simulation model), simplified version or progress indicator of the batch queries being executed; also for preparing the data for a fast visualization in the visualization client.

Further details of the architecture can be found in deliverable D2.4 and D2.5.

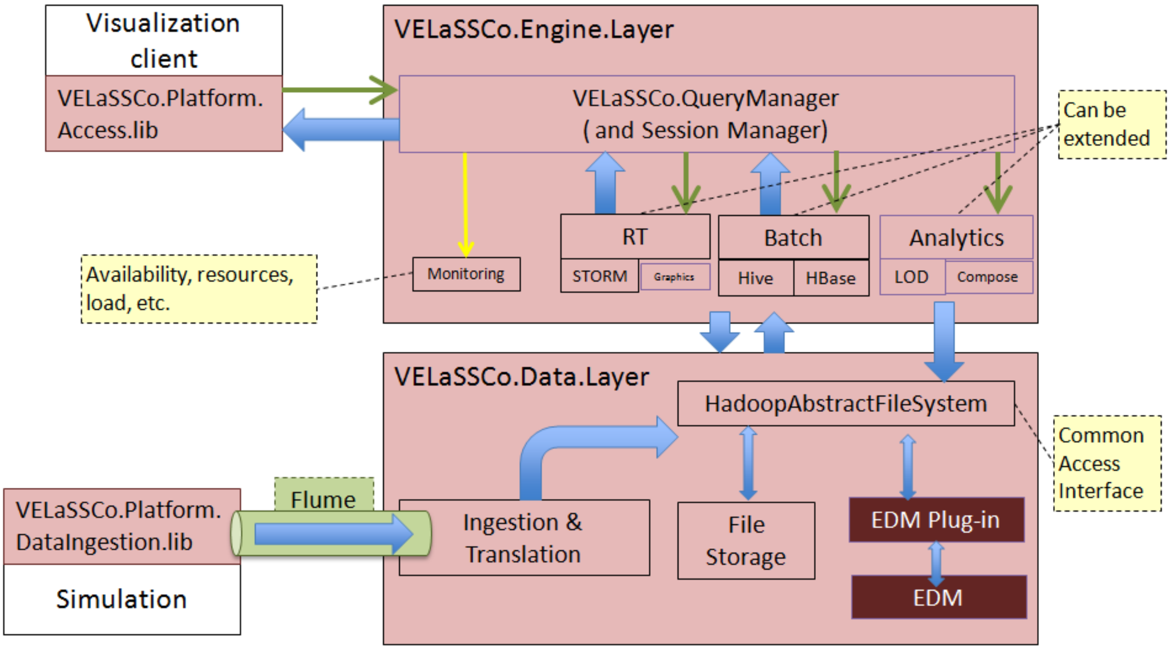

Figure 9. *Layers and modules involved in the platform (November 2014)*

Figure 10 shows the definitive version of the architecture of VELaSSCo Platform, after several iterations done within the project lifetime.

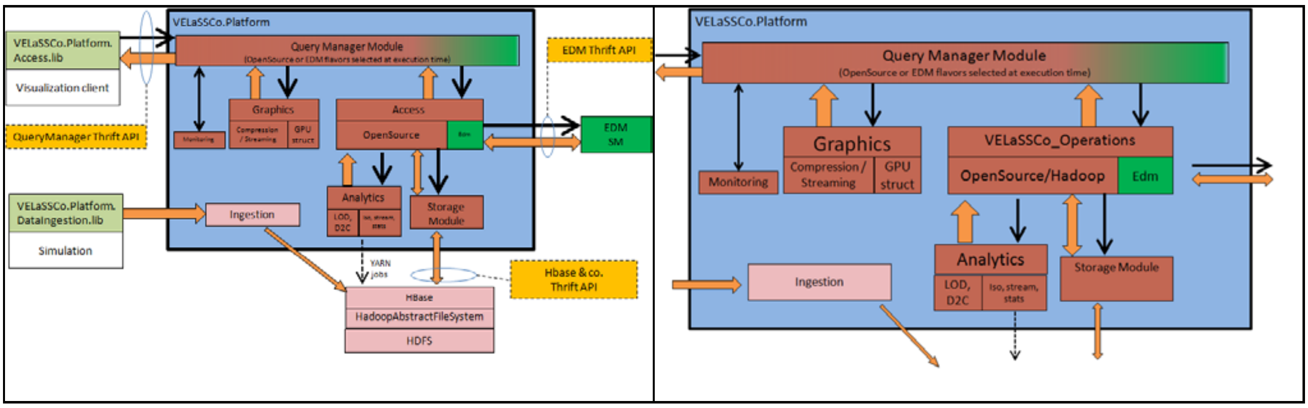

Figure 10. *Definitive version of the architecture of VELaSSCo Platform (left), and a close-up view of the VELaSSCo Query Manager (right).*

Extension to cloud

Although VELaSSCo Platform prototype has been implemented in an HPC infrastructure, big data architectures are mostly designed to work on cloud, so it is natural to think about porting it to cloud. Having this in mind, an extensive analysis of the existing cloud solutions has been made in the deliverable D2.3 – *HPC cloud infrastructure specification document suitable to the needs of e-*

Science, as well as the main characteristics of HPC, Cloud and HPC-Cloud infrastructures, including their specifications and requirements suitable to affect the VELaSSCo Platform architecture. The main conclusion is that, from a technical point of view, the architecture designed for the platform should be portable on all these infrastructures with only slight modifications.

One of the important aspects to be considered when porting an application to Cloud is the business model. This may not be a technical issue, but the new paradigms of business models raised up together with the cloud solutions has a deep impact in the exploitation strategies. IaaS (Infrastructure as a Service), SaaS (Software as a Service) and other related business models may affect to the decision of porting VELaSSCo platform to cloud in the near future, knowing that its architecture has been designed to be portable (technically).

1.3.4 VELaSSCo Platform development and implementation

Data Storage

The Open-source option of the platform has been implemented in Hadoop's framework. Particularly, Apache's HBase has been chosen to store engineering simulation data in a persistence layer. HBase is a fast and efficient solution to store and access data with a fast indexing strategy for tabular data, it is tightly integrated with HDFS and several open standard tools such as Apache Kafka, Apache Flume, Apache Spark, and Impala are designed to be on top of HBase.

In a nutshell, the platform uses three tables: one with the list of stored simulations models and a summary of their contents; a second one with metadata information relative to the meshes and results present in the simulations, like the names, types and sizes of the meshes and results; and a third one to store the 'bulk-data', the elements and results of the simulation model. Details about the distribution and structure of the data can be found in deliverables D2.6 and D2.7.

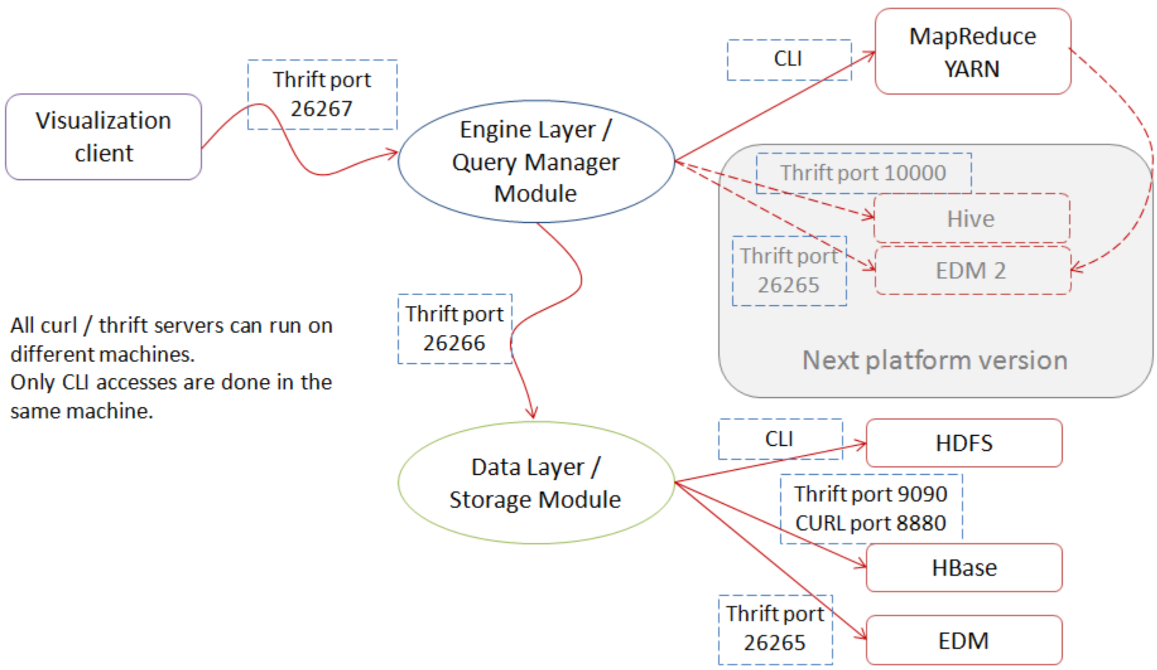

Figure 11. Processes diagram of the VELaSSCo platform of the 1st Evaluation Event

Injection

The ingestion module uses Flume agents to read the data files, generate injection events, send them to the HBase server, then process them and inject the data in the HBase tables.

First prototype of the Platform

The first version of the platform, evaluated in the first evaluation event, basically follows the architecture depicted in Figure 9, with small differences, like the added ‘Graphics’ module, removal of the ‘Batch’ and ‘RT’ modules, and embedding more complexity in the Data Layer. Each layer was implemented as stand-alone programs, which communicate using the Thrift communication protocol. A diagram of the processes involved and their protocols are shown in Figure 11.

Final version of the Platform

The final version of the platform, evaluated in the last evaluation event, follows the architecture shown in Figure 10, which basically merges the Data and Engine Layer and simplifies the workflows of the VQueries. This version of the architecture has one disadvantage though, it makes necessary to duplicate the analytics and graphics modules for both versions of the platform: the open source and EDM’s version.

Development plan

To allow a collaborative development and implementation of the platform, some conceptual entities were defined connecting the architecture (conceptual) and the implementation (coding). These entities have been the base for the implementation plan followed during the project (more details in deliverable D3.1).

The platform is divided into three conceptual Layers:

- **External Layer**: related to the injection of simulation data and the communication between the visualization client and the VELaSSCo platform;
- **Data Layer**: related to the organization and storage of the simulation data;
- **Engine Layer**: related to the management, execution of the VELaSSCo queries, which may include analysis and processing of the simulation data and ensure the continuous feed-back of information to the visualization client by using multi-resolution models, compression and other GPU friendly techniques.

Later the Data Layer and Engine Layer have been merged to be the VELaSSCo platform, which encloses the Query Manager Module, the Graphics, Analytics, Data Access modules.

The workflow of the functionalities available in the platform is organized in **VQueries**. A VQuery is the minimal functional entity of the platform, and represents the entire workflow needed to perform the requested action since the visualization client launches the query (upon user request) until it receives the result from the platform (eventually render the result). As an example, when the final user wants to visualize a volume mesh, the visualization client will issue a GetBoundaryMesh VQuery to the platform.

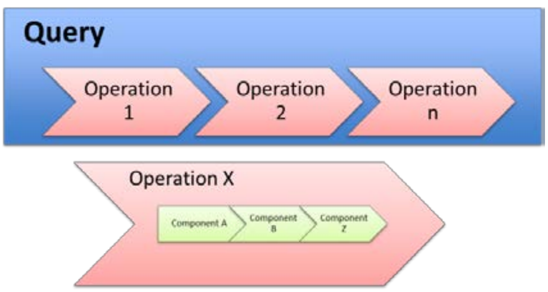

Figure 12: Structure of a VQuery and Operations

In order to organize the implementation plan, and to avoid the repetition of code, each VQuery has broken down into different **Operations** organized in a specific workflow (see *Figure 12*), and each operation is formed by different **Components** (which are the parts containing the code itself: functions, etc…).

All the VQueries have a common part, which is presented hereafter (*Figure 13*), so once all the common operations are implemented, the development and implementation of a new VQuery can be done relatively independently on the platform. This modularity has allowed a collaborative programing within the project among all the partners of the consortium, optimizing the efforts.

Deliverable D3.1 contains all the details of this implementation plan, and also the forms defining all the VQueries and Operations involved in VELaSSCo Platform (description, input/output data, dependencies, etc…).

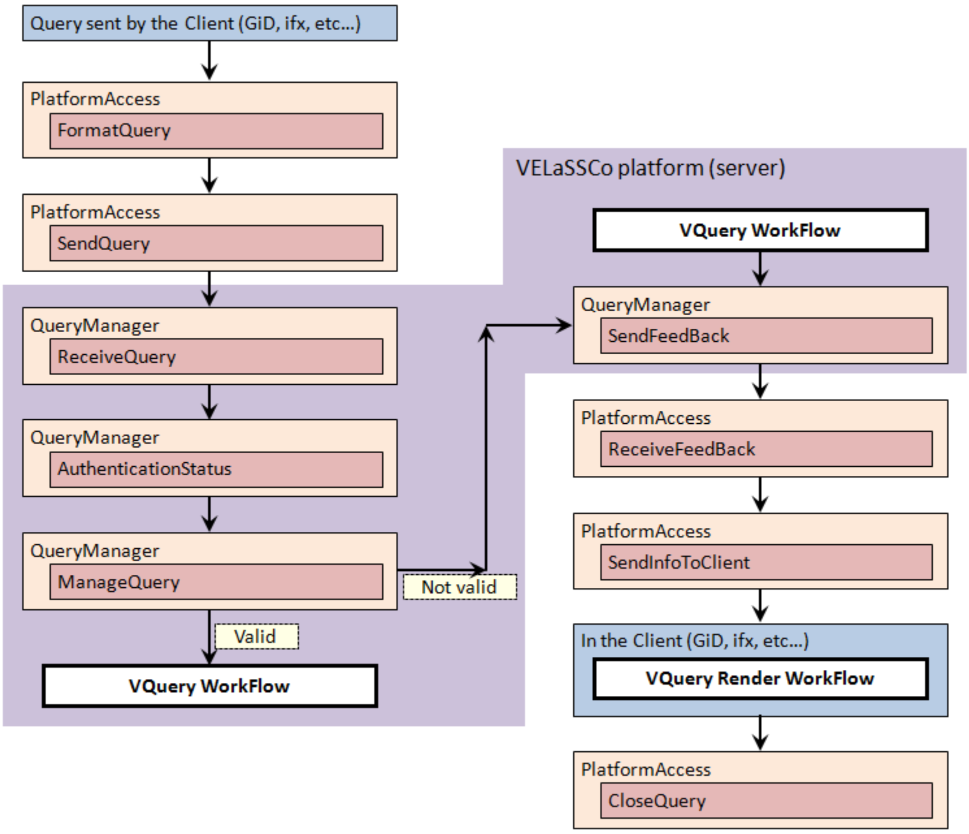

Figure 13: Common part in all VQueries workflow

VQueries implemented

As mentioned earlier a VQuery is the minimal functional entity of the platform which may be called from the visualization client or the injection process.

As described in deliverable D2.5, the triggered Vqueries can be twofold: simple and complex. A simple query directly retrieves information from the storage layer, like GetMeshVertices, OpenModel or UserLogin; while complex queries produce content from a computation, like GetBoundaryMesh, GetSimplifiedMesh or GetDiscrete2ContinuumTransformation.

D2.2 outlined two types of workflows: simple and complex. Simple workflows are the execution of the VQueries on-demand. The user interacts with the visualization client, which in turn issues one or several VQueries and waits for its results. The VELaSSCo platform then executes the VQuery and returns its result. Three kinds of complex workflows were identified: “Store calculated data for late re-use”, “Navigational and streaming”, when the user travels along the model, a streaming of incoming data is being generated; or “Multi-resolution” when a low-resolution version of the results are shown to the user while the full-resolution is being calculated.

All implemented VQueries implements this simple workflow and some complex VQueries implement the “Store calculated data for later re-use” complex workflow, a list of these can be found in deliverable D3.7. Some of the implemented VQueries are fundamental pieces to implement the “Multi-resolution” complex workflow.

The final version of the platform offers 32 VQueries. Here is the list grouped by functionalities:
Session Vqueries:

- UserLogin, UserLogout
- GetListOfModels, OpenModel, CloseModel

Direct Result Vqueries:

- GetListOfMeshes, GetMeshDrawData, GetMeshVertices,
- GetListOfAnalysis, GetListOfTimeSteps, GetListOfResults, GetResultFromVerticesID

ResultAnalysis Vqueries:

- GetBoundingBox, DeleteBoundingBox,
- GetBoundaryOfAMesh, DeleteBoundaryOfAMesh
- GetSimplifiedMesh, DeleteSimplifiedMesh
- GetDiscrete2Continuum
- DoStreamLinesWithResults
- GetIsoSurface
- GetSimplifiedMeshWithResults, DeleteSimplifiedMeshWithResults
- DeleteAllCalculationsForThisModel
- GetVolumeLRSplineFromBoundingBox, DeleteVolumeLRSplineFromBoundingBox

Monitoring Vqueries:

- GetStatusDB, GetConfiguration, SetConfiguration, ErrorMessage, StartTestServer

Data Ingestion Vquery:

- IngestDataFiles

Data analytics for particle-based simulations

As commented previously, VELaSSCo Platform deals with results from FEM and DEM simulations. Obviously, the platform can process, query and visualize the particles and their information separately, but one of the main functionalities of VELaSSCo Platform when dealing with DEM simulations, is the so called discrete-2-continuum transformation, which allows the temporal and spatial averaging of the discrete data. By means of the spatial averaging, the information from the DEM simulation data (particles and contacts) is projected onto a continuum field (more details in deliverable D3.6).

The outputs of this methodology are new results that are related to the intermediate length scales of the simulated system such as bulk scale. These results are defined on a FEM-like mesh, so once the discrete-2-continuum transformation is performed, all the post-processing and visualization techniques available for FEM simulations can be applied to it. Some examples of results visualization on the discrete-2-continuum transformation of a DEM simulation are depicted in Figure 14 and Figure 15.

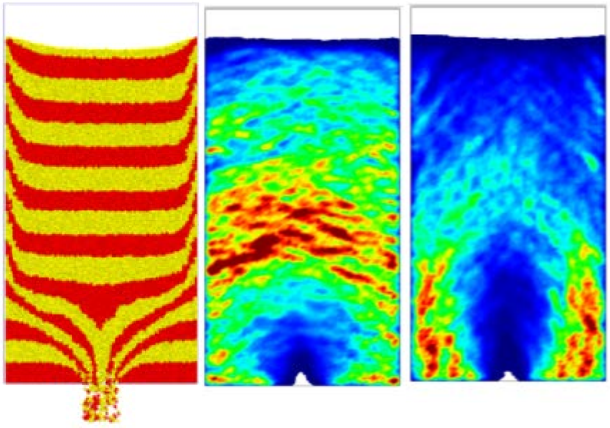

Figure 14: *Example of spatial averaging to compute the contact stress from the contacts data of a DEM simulation of a silo discharge: (a) Snapshot of the particles; (b) Sxx component of the contact stress tensor (lateral stress); (c) Szz component of the contact stress tensor (vertical stress).*

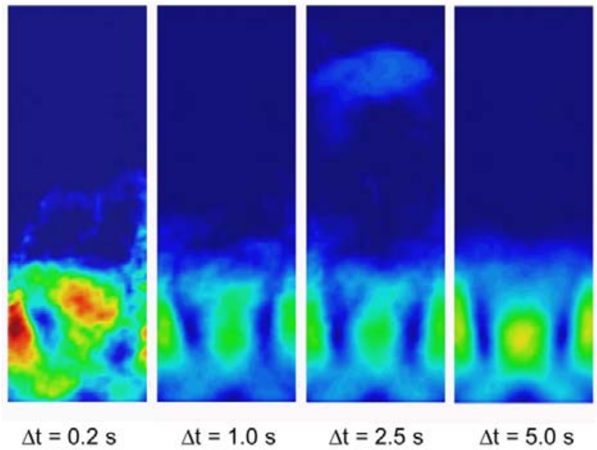

Figure 15: *Example of the effect of temporal averaging at different time scales for the DEM simulation of a fluidized bed. The different snapshots are obtained from the same DEM dataset but varying the width of the temporal window (Δt) to conduct the averaging.*

Data Simplification algorithms

As commented previously, a smart way of visualize huge amounts of data is precisely not to use that huge amounts of data, but a simplified set of them. One simple example is that if some set of triangles must be painted in the screen, independently on the number of triangles present in the original data, there will never be more than the available pixels to be used.

In that direction, one of the ways VELaSSCo is tackling the problem of huge amounts of data is to simplify these data a priori, so as when the user query on that data, or wants to visualize it, a smaller set of data is used. This improves the performance in several aspects: post-processing operations, transferring data from the platform to the visualization client, and managing it in the visualization client.

In the frame of VELaSSCo Platform we have developed two main strategies for simplification of data: mesh simplification, and Locally Refined (LR)-BSplines representation.

The **mesh simplification** strategy followed is based on a vertex clustering of the origin mesh, considering both the geometrical features of the mesh and the results onto it, so the level of approximation from the origin to the final model (mesh+results) can be adapted accordingly to the user or hardware needs. The more simplification, the less the reduced model occupies, and the less accurate is the representation of the model. An example of this simplification of the mesh is depicted in Figure 16.

LR-BSplines are used as a new class of data analytics for FEA and DEM. For fields with an overall smooth behaviour, a drastic reduction in data volume is obtained, leading to the capability to query and interact with the model in real-time, even from extremely large origin data sets (details on this can be found in deliverable D3.7. The visualization capabilities implemented for this LR-BSplines structures include dynamic selection of streamlines, iso-surfaces, volume transparencies, among others. Some views of this kind of visualizations are depicted in Figure 17.

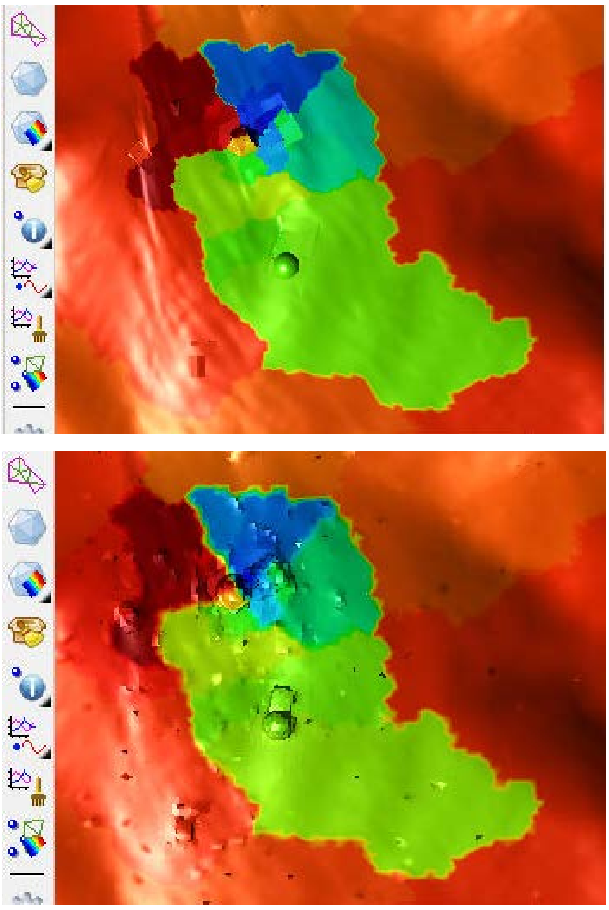

Figure 16: *Partition Index result on the boundary mesh of original Telescope model, in the image below the Partition Index result on the simplified volume mesh resulting from the* GetSimplifiedMeshWithResults *VQuery. The simplified volume mesh (7% of the size of the original) is transferred to the visualization client for fast analysis and visualization.*

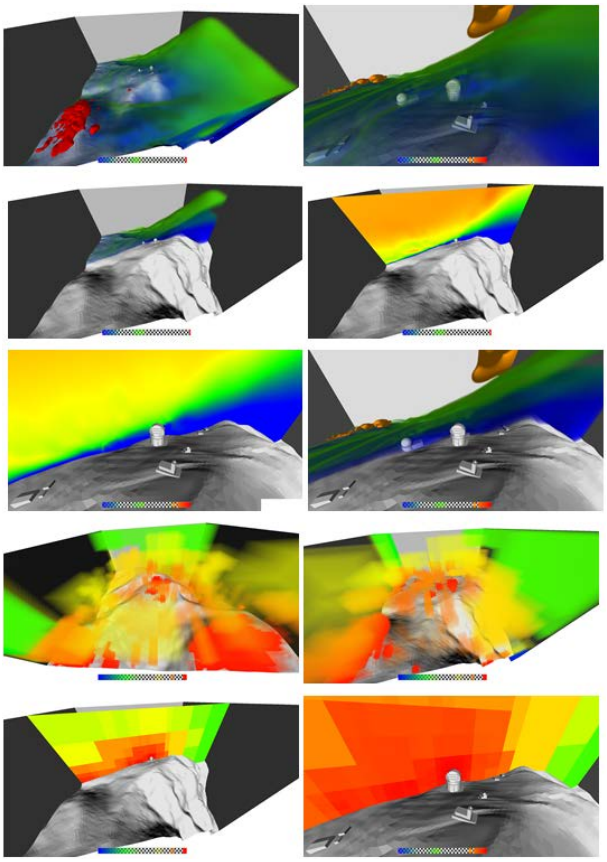

Figure 17*: Visualization of the volumetric LR-BSpline model. Top: Volumetric scalar field with isosurfaces (right: details around the telescope). Second row: Clipping and cut planes, and third row: details of these close to the telescope. Bottom two rows: Visualization of the LR mesh, colour-coded by size, from smallest (red) to largest (blue).*

Visualization plug-ins

VELaSSCo Platform has been designed considering that any kind of visualization client can use it, of course, following the specifications of the access library implemented. In the frame of the project, two visualization clients has been used: **iFX** (from Fraunhofer) and **GiD** (www.gidhome.com, from CIMNE).

For each visualization client, a specific plug-in has been implemented in order to adapt it to the communication with the VELaSSCo Plaform, and to include the required options at GUI level to use properly the functionalities of the platform.

Some snapshots of iFX and GiD visualization clients are depicted in Figure 18 and Figure 19.

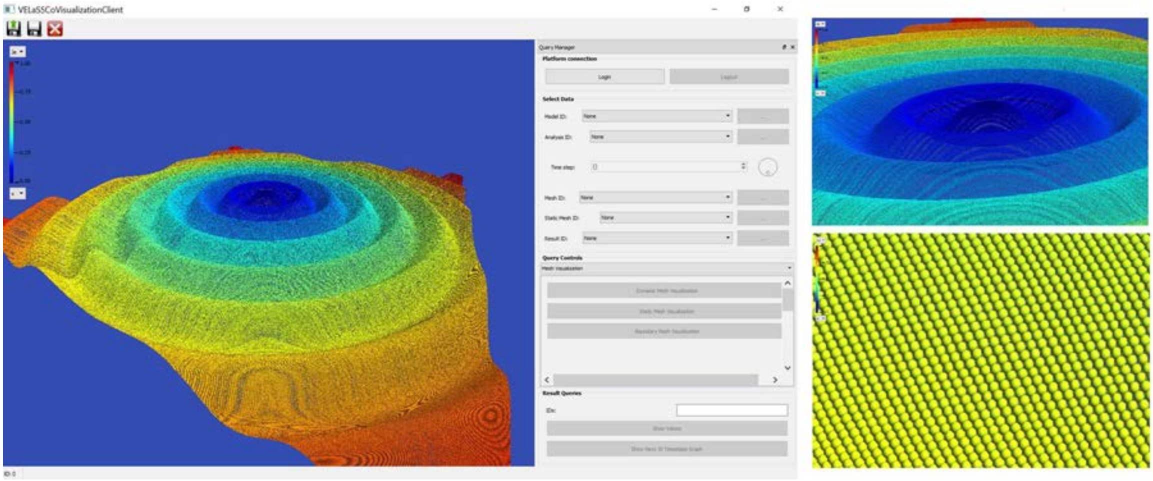

Figure 18: *Fraunhofer Visualization Client. More than 23 Million particles can be rendered pixel-accurate and shaded with more than 10 frames per second. In addition with new improvements, each particle individually can be picked. Interacting with particle (e.g. pick and rotate around one of them) feature is interactive now.*

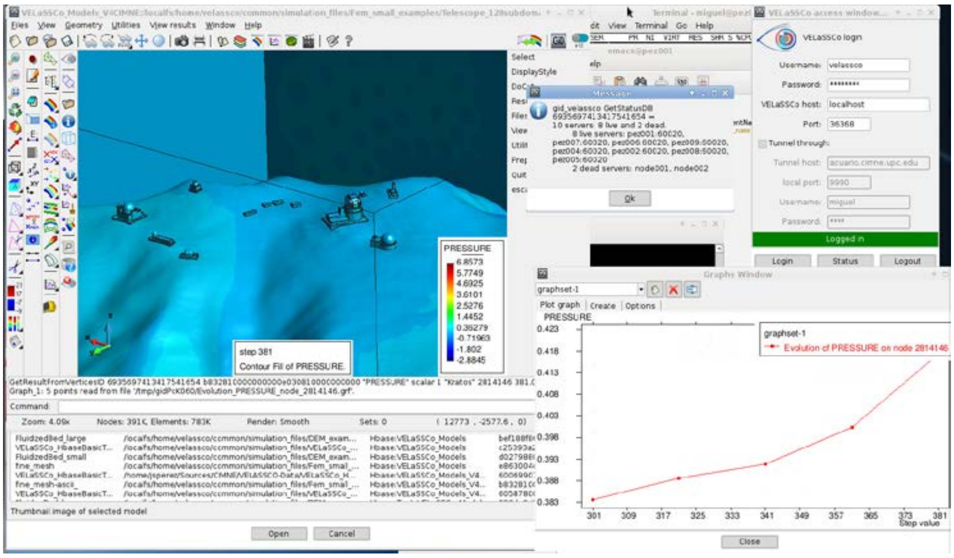

Figure 19: *GiD visualization client with the developed VELaSSCo plug-in.*

Advanced graphics technology

The latest GPU techniques have been used to visualize and render the LRBSplines representation of the volume data as shown in Figure 20.

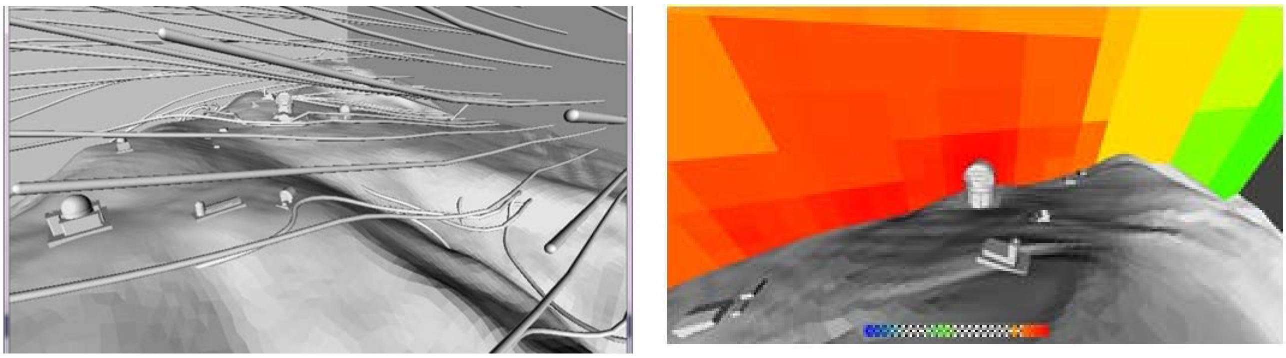

Figure 20: SINTEF's LRBspline visualization plug-in in action.

1.3.5 Evaluation of the Platform

From the beginning of the project, it has been clear that the platform to be developed should be tested and evaluated by the members of the user panel, in order to get a realistic feed-back about it. For this purpose, a solid methodology for system evaluation has been adopted, and two evaluation events have been built allowing the user panel to use and evaluate the platform.

Evaluation methodology

The Goal/ Question/Metric (GQM) [3] methodology has been selected as the core methodology for the system evaluation. GQM is about identifying business goals to drive the characterization of the right metrics, and finally gathering the measurement data and making effective use of the measurement results to drive evaluation and improvements of the system. GQM is a contrasted methodology used now for decades in industry for system evaluation and improvement. In VELaSSCo, GQM helps to categorize the main business goals, taken from the User Panel and knowledge from the partners. From the goals, the methodology identifies in a systematic manner the right questions to enable the technical requirements assessment of the VELaSSCo framework, identifying the main features the project is aiming at. GQM was chosen precisely to get from the objectives to the actual metrics, providing a measurement model on three levels [5]:

- Conceptual level (goal). A goal is defined for an object, for a variety of reasons, with respect to various models of quality, from various points of view, and relative to a particular environment.
- Operational level (question). A set of questions is used to characterize the assessment/achievement [how] of a specific goal going to be performed is based on some model.
- Quantitative level (metric). A set of data is associated with every question in order to answer it in a quantitative way.

VELaSSCo has defined the user and technical requirements following the GQM methodology. The evaluation plan is based in applying iteratively the GQM methodology in cycles, where each cycle refines the measurement program and the VELaSSCo system.

Figure 21: *VELaSSCo GQM standard cycle*

The instantiation of single steps in GQM cycle for VELaSSCo (depicted in Figure 21) comprises the following actions (detailed in deliverable D1.5):

- Identify GQM Goals.
- Develop GQM Plan.
- Derive Measurement Plan.
- Data Collection.
- Interpret Collected Data.

The proposed framework has five key dimensions (displayed in Figure 22) to describe the VELaSSCo System: End-User Functionalities, SW Architecture and Deployment Environment, Navigation and Interaction, Algorithms and Views.

End-User Functionalities refer to business actions offered by VELaSSCo system from the end user perspective. **SW Architecture** and **Deployment Environment** consider the definition, implementation and deployment in the production environment of the SW pieces integrated by means of an integration framework. **Navigation and Interaction** characterize the ease of use of VELaSSCo SW System. **Algorithms** represent the processes of multi-resolution, coarsening, coordinates, connectivity and results compaction. Finally, **Views** characterize the perspective of the VELaSSCo observers focusing on effectiveness aspects that can help to these to take decisions.

The GQM paradigm will be used by each of these five key dimensions to identify the questions and to then enable the formation of the framework features.

Figure 22: *VELaSSCo evaluation framework dimensions*

Following the steps of the GQM evaluation methodology proposed, the evaluation team will set up a series of metrics to evaluate the system. In many cases, metrics can be scenario or use case-specific. This means that the system would need to achieve different quantitative measurements for the same metric, or even some metrics would not apply to certain scenarios.

More details on how GQM methodology has been applied to the VELaSSCo Platform evaluation can be found in deliverable D1.5.

SERVQual methodology

Although the GQM methodology is well stablished for evaluating software components, and it follows a well known standards, after a recommendation from the reviewers of the project in the second review meeting, the consortium decided to use complementary to GQM a simpler methodology, trying to focus more in the strong points of VELaSSCo Platform. For this purpose, the

SERVQual methodology has been adopted (in the final evaluation event) for the evaluation of the usability of the platform.

Measuring the quality of a service can be a very difficult exercise. Unlike product where there are specific specifications such as length, depth, width, weight, colour, and so on, a service can have numerous intangible or qualitative specifications. In addition, there is there expectation of the customer with regards the service, which can vary considerably based on a range of factors such as prior experience, personal needs and what other people may have told them.

SERVQUAL represents service quality as the discrepancy between a customer's expectations for a service offering and the customer's perceptions of the service received, requiring respondents to answer questions about both their expectations and their perceptions. The difference between expectations and perceptions is called the gap which is the determinant of customers' perception of service quality.

Figure 23 shows the adaptation of SRVQUAL methodology into VELaSCCo project considering Architecture, Usability and Effectiveness dimensions.

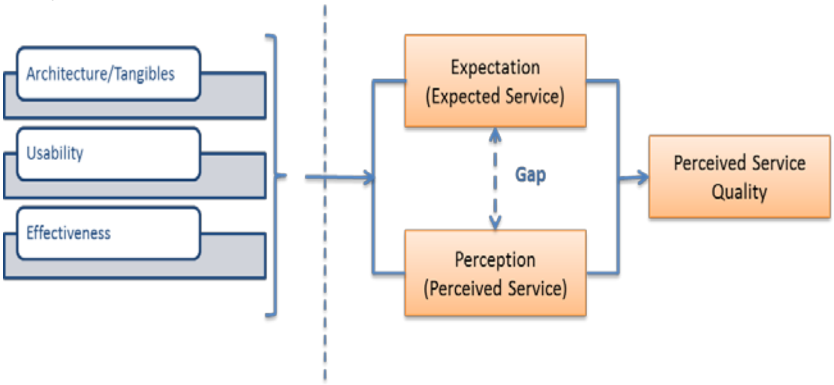

Figure 23. *SERVQUAL methodology for VELaSSCo project.*

Evaluation Events

Two evaluation events were performed: one in **February 2016** with the first prototype of the VELaSSCo Platform, and another one on **November 2016** with the final prototype of it.
The scenarios used in both evaluation events involved DEM and FEM simulations, with the following specific characteristics:

- **DEM**: The VELaSSCo platform should handle three groups of scenario:
  - One big model with few time steps.
  - One small model with many time steps.
- **FEM**: The VELaSSCo platform should handle two possible scenarios:
  - Huge simulation with distributed results and several time steps of an HPC cluster.

For the **first evaluation event**, only a few VQueries were implemented in the platform, nevertheless it was useful to detect some bottlenecks and focus the developments to be done during the third year of the project. This evaluation event was held in parallel in Barcelona (CIMNE's premises) and Edinburg (UEDIN's premises), and was organized dividing the users in two groups, one for each visualization client available (GiD and iFX).

In both cases, facilitators from the consortium assist the member of the user panel to follow the evaluation event properly, and observe their behaviour. A picture of some of the attendees to the evaluation event in CIMNE (Barcelona) and the facilitators for it is shown in Figure 24.

Figure 24*: Facilitators observing user behaviour at the first evaluation event in CIMNE*

The Platform was deployed in Acuario cluster (see 0, Figure 7), which was temporary provided by CIMNE because of the delay of EDDIE cluster access to the consortium. Details on this evaluation event can be found in deliverable D1.6.

Besides the facilitators' feedback, in the scope of these evaluation event the following tools were used to gather evidence:

- Nagios[1] is an open source software monitor tool which enables monitoring your entire IT infrastructure to ensure systems, applications, services, and business processes are functioning properly. Besides the obvious use of Nagios to monitor cluster resources in order to detect any failure of servers, network connection, etc., we used it to obtain performance metrics valuable for evaluation purposes.
- VELaSSCo Query Manager and Access Library Logs: As main mediators' modules between the visualization client and the VELaSSCo platform, partners in charge of the implementation of these two components were instructed to add a few lines of code to actually provide timely statistics about the usage of the application during the evaluation events. These logs are provided for further analysis.
- Usability Questionnaire measuring the user-friendliness and appropriateness of the solution.
- Effectiveness Questionnaires to measure the performance of the visual and real time aspects based on the performance indicators (defined in deliverable D5.1).

Besides these tools for data collection, the SOASTA CloudTest Lite tool has been considered and used in the scope of the evaluation performed in the Acuario cluster in order to simulate a number of concurrent users of the platform. Details on this tool can be found in the deliverable D1.6.

The evaluation event consisted in following the steps of some predefined use cases, involving the different VQueries implemented in the platform, applied to two specific models (one with FEM and the other with DEM data).

The FEM model (Telescope) comes from a simulation of the airflow around a great telescope building, considering also the terrain, which was run on 128 cores. The model has a total of 4 M nodes and 28 M tetrahedral, and 19 time-steps with Partition Index, Pressure and Velocity vector as results. A view of the Telescope model is depicted in Figure 25.

The DEM model (Fluidized Bed) comes from a simulation of a very common process in pharmaceutical and chemical industries, which is the fluidized bed. The characteristics of the data are:

[1] https://www.nagios.org/

- Fluidized bed process very common in pharmaceutical and chemical industries.
- ~ 12000 particles / time-step
- ~ 3000 contacts (p2p and p2w) / time-step.
- 40 000 time-steps.
- Particle results: mass, volume, velocity vector
- Contact results: Force vector.

A view of the Fluidized Bed model is depicted in Figure 26.

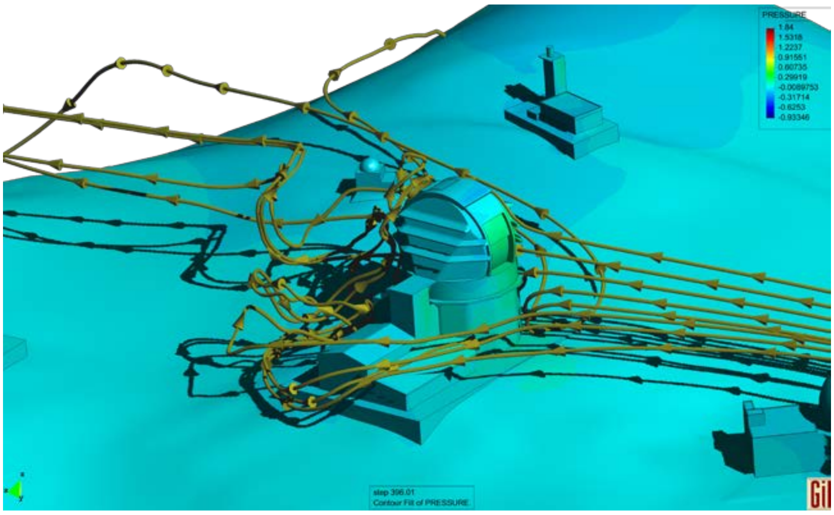

Figure 25*: View of the Telescope model, used for the FEM use cases in the first evaluation event*

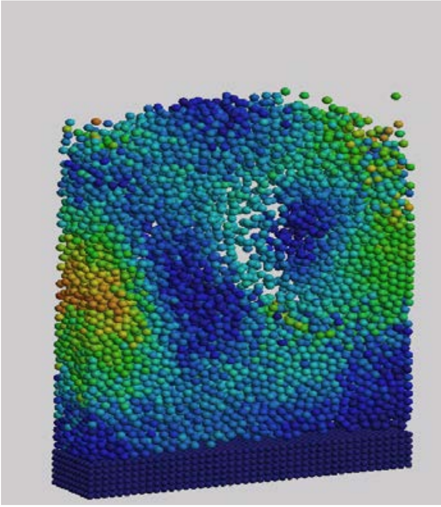

Figure 26: *View of the Fluidized Bed model, used for the DEM use cases in the first evaluation event*

For the **final evaluation event**, the final prototype version of VELaSSCo platform was used, deployed in EDDIE cluster (see 0, Figure 6). Details on the evaluation of the platform in this evaluation event can be found in deliverables D5.3, D5.5, D5.7 and D5.9. In this evaluation event, the SRVQual methodology was used for the Usability evaluation.

In this case, the event was done telematically, organized in the following way:

- Two days were fixed to hold a webex meeting with the user panel members to explain the use cases. One day for the use cases using GiD as visualization client, and the other devoted to iFX. This second meeting was recorded, so as any member of the user panel could access to the information if it was not possible for him/her to attend the meeting.
- All the documentation related to the use cases was uploaded in the website of the project, as well as the connection details to EDDIE cluster, to be able to connect the platform. In this occasion, previous to the evaluation event, each member of the user panel willing to participate was contacted in order to provide the confidential data to log-in in EDDIE cluster.
- Then, one week was left for the users to follow the use cases remotely, and fill-up the SRVQual questionnaire remotely.
- After this week, all the data was processed in order to obtain an objective measure of the evaluation of the platform.

For this final evaluation event, larger models were used to evaluate the Platform, as the Platform was more mature, and it was deployed in EDDIE cluster (with access to more resources).

The FEM model was a model of Barcelona city, where the airflow around it considering the buildings (in a simplified version) and the Digital Model of the Terrain was considered. Some view of the model is depicted in Figure 27. The simulation providing these results was run on 3072 cores (using Kratos Multiphysics solver), and the data was a mesh of **60 Millions** of nodes and **340 Millions** of tetrahedra divided in 3072 partitions, and 240 time steps with Partition index, Pressure and Velocity vector as results for each node. To have an order of magnitude of the size of this model, the results file of this simulation written by the solver occupies around **2TB** in ASCII format.

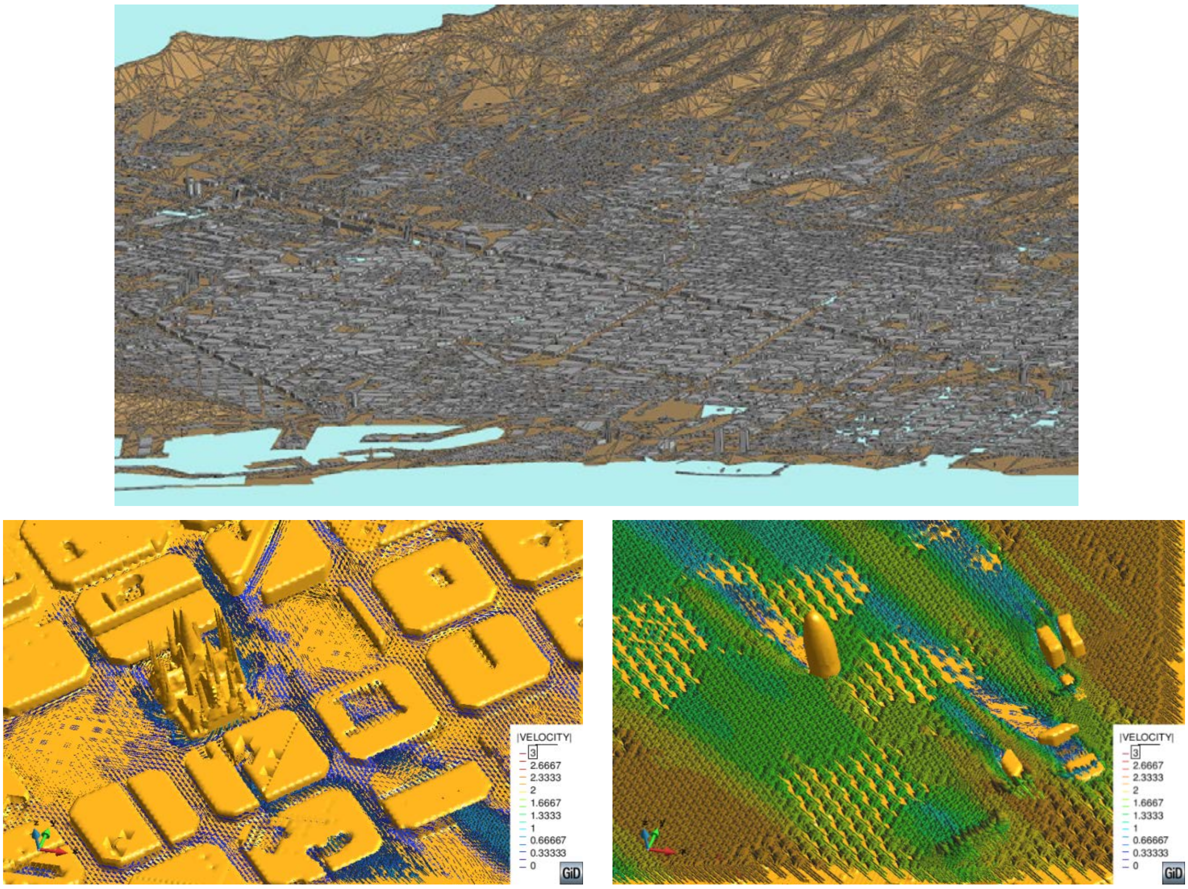

Figure 27: *Views of the Barcelona model used for the FEM use cases in the final evaluation event.*

Two models were used for the DEM use cases in the final evaluation event: Rail Embankment and Fluidized Bed Large.

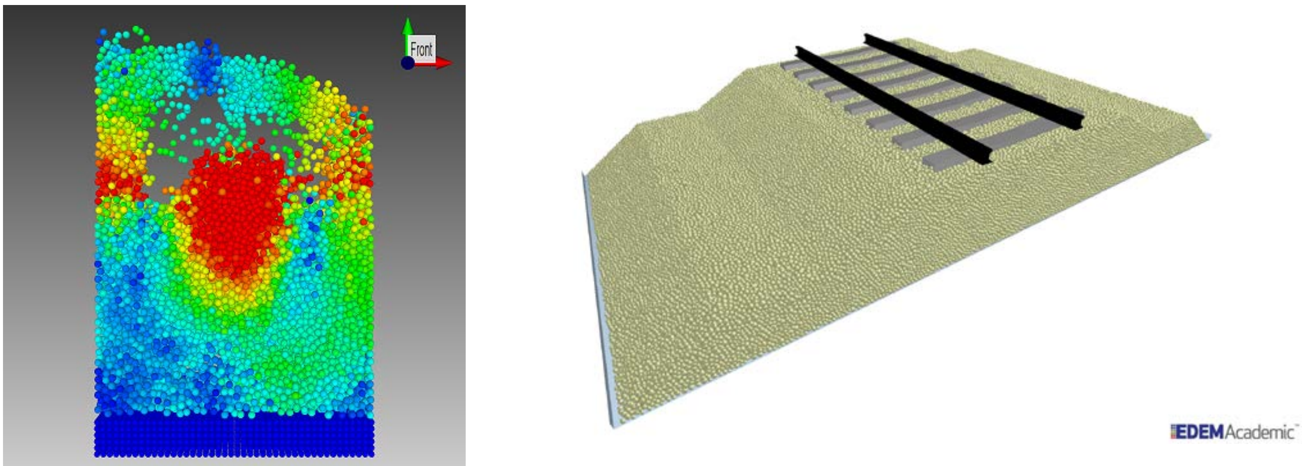

Figure 28*: Views of Fluidized Bed Large (left) and Rail embankment models, used for the DEM use cases in the final evaluation event.*

A view of them can be seen in Figure 28. This Fluidized Bed model is larger than the used in the first evaluation event: 12K particles per time step, 3K contacts (p2p and p2w) per time step, 48K time steps, mass, volume and velocity vector as particle results, and force vector as contact result.

The characteristics of the rail embankment model are: 207K particles per time step, 560K contacts (p2p and p2w) per time step, 9102 time steps, and mass, volume and velocity vector as particle results, and force vector as contact result.

Results from evaluation

The data collected after the celebration of first evaluation provided information about basic functionalities aspects of VELaSSCo platform developed so far. At that stage of the project, the more advanced modules developed were those related with database storage in a proper data model, simulation data injection and client visualization tools. The main challenge so far was the proper interaction among these services to allow complete workflow execution successfully.

Our analysis over the results obtained about the features evaluated was satisfactory for most of them. Finally a set of improvements were proposed in order to improve the VELaSSCo platform and visualization clients' plug-ins in future releases and ensuring compliance with all the objectives set out in the tables GQM.

Regarding to final evaluation event it is important to mention that after analyzing the data to assess conformance to the goals the conclusion was that most of objectives were achieved. A checking process about the coverage of improvements proposed after first evaluation event was performed for all the evaluation dimensions. The specific actions and status of them are listed below:

Improvement proposed Y2	Status after final evaluation
VELaSSCo Injection Module should be able to work properly with **Extra Large** simulations (hundreds of millions of data events generated) maintained the current throughput. To get this, the deployment of flume agent's topologies will be needed.	Achieved
From the Simulation Configuration point of view, VELaSSCo platform should be able to work properly with:	
• **Extra Large** simulations (>1 TB)	Achieved
• Simulations with a **Large** size (millions of particles) in terms of number of particles	Achieved
• Simulations with a **Large** size (millions of contacts) in terms of number of contacts	Achieved
• Simulations with a **Large** number of results at particle level (> 10)	Not Achieved
• Simulations with an **Large** number of results at contact level (> 10)	Not Achieved
Provide a VPN connection to EDDIE cluster agile and reliable supporting multisession for the "VELaSSCo" user.	Achieved
Improve significantly the performance on critical Vqueries as VQ-214 (GetBoundaryOfAMesh), VQ-116 (GetMeshVertices), and VQ-200	Achieved

(GetBoundingBox). It will be needed to reformulate the query on HBase in order to avoid scan process as part of query execution. Other option to explore will be to investigate how Thrift is handling the transfer of complex data structures.	
VELaSSCo platform has currently a **Low Analyzability** degree that should be increased by means of:	
• Improving Audit Capability	**Not Achieved**
• Including a significant number Diagnostic functions	**Not Achieved**
• Providing new monitoring services and alerts in Nagios in order to increase for instance the Failure analysis capability as well as the Monitoring capability for specific operations in all SW modules that VELaSSCo platform is composed by.	**Achieved**
VELaSSCo platform should increase the Testability level providing for each one of the SW modules a simple use case test battery in order to an end user and maintainer easily perform operational testing without additional test facility preparation.	**Partially Achieved**

Finally, three important questions were included in the evaluation questionnaires in order to stablish some conclusions related to overall satisfaction degree of the user panel members:

1. Would you use VELaSSCo Plugins?
2. Would you buy VELaSSCo Plugin?
3. Would you recommend VELaSSCo Plugin?

The overall score was slightly up to 3 in a five-point Likert-style scale from 1=very unlikely to 5=very likely.

Some performance numbers

To conclude with the evaluation part, it is interesting to include in this document some internal evaluation measures done by the consortium, considering the very last implementations of the VQueries. Detailed information on this can be found in deliverable D3.7. Data sizes of the different test models used are shown in Table 1.

	# of elements	# steps	ASCII file sizes	HBase table size
Telescope model	24 M	19	8.4 GB	39 GB
Barcelona 8m model	100 M	240	526 GB	2.4 TB
Barcelona 4m model	340 M	264	2.0 TB	9.3 TB

Table 1: Size of the FEM models used.

Boundary of a mesh

The performance of the GetBoundaryOfAMesh (VQuery in charge of extracting the triangles boundary of a volume tetrahedral mesh) can be observed in Table 2, which shows the executions times for the Telescope, Barcelona small and Barcelona big models in UEDIN's EDDIE cluster:

	Telescope	Barcelona 8m	Barcelona 4m
# splits	40	181	268

	# vertices	4.5 million	18 million	59 million
	# elements	24 million	99 million	340 million
	Task			
GetBoundary	MR job	22.18 s.	75.26 s.	2m. 31s.
	GetVertices	20 s.	90 s.	5m. 21s.
	Other			49 s.
	Total	87 s.	208 s.	8m. 41s.

Table 2: Execution times for the GetBoundaryOfAMesh YARN job on EDDIE cluster

Mesh simplification

The performance of GetSimplifiedMesh VQuery is shown in *Table **3***.

	Task	Telescope	BCN 8m 100M	BCN 4m 340M
Grid of 9^3 cells		375v 1873e	36v 22e	55v 16e
	MR job	12,67 s.	61.64 s.	93,89 s.
	Java	52.37 s.	90.04 s.	131.96 s.
	Total	64 s.	100 s.	138 s.
Grid of 247^3 cells		305kv 1.62Me	113Kv 629Ke	89Kv 393Ke
	MR job	26.20 s.	61.13 s.	88.52 s.
	Java	54.98 s.	89.03 s.	116.8 s.
	Total	75 s.	114 s.	125 s.
Grid of 1023^3 cells		801Kv 4.74Me	1.68Mv 9.85Me	1.36Mv 7.07Me
	MR job	41.93 s.	113.4 s.	123,11 s.
	Java	70.53 s.	141.89 s.	179.48 s.
	Total	113 s.	197 s.	212 s.

*Table **3**: Execution times for the GetSimplifiedMesh on EDDIE cluster*

LR-BSplines

The following measure refers to the performance of the CreateVolumeLRSplineFromBoundingBox VQuery. It was taken at a desktop computer for the Telescope model. The implemented version is only partially parallelized and we should thus expect a modest speedup when testing the performance on EDDIE. We have let the algorithm choose a suitable error tolerance based on a standard deviation-measure for the model.

Model	Input Size	Output Size	Exec. Time
Telescope	192 MB	2.2 MB	83 min

Discrete-2-continuum

The results of GetDiscrete2ContinuumOfAModel VQuery applied to models shown in *Table **4*** are presented in *Table **5***.

Model	Size	Particles/Contacts	Total timesteps
Fluidised Bed	40GB	11880/3000	48000
Railway Embankment	550GB	207K/500K	9000

*Table **4**: DEM models used to evaluate discrete-2-continuum VQuery*

Run	Execution time
Fluidised bed/ 1timestep	0m23
Fluidised bed/ 1000 timesteps	1m50
Railway embankment/ 1 timestep	1m26
Railway embankment/ 10 timesteps	2m19

*Table **5**: Performance of Discrete-2-continuum VQuery run in EDDIE*

EDM version

Execution times for some of the queries when executed on three different configurations in EDDIE are shown in *Table **6*** (using 1, 8 and 27 HPC nodes).

Number of HPC nodes	1	8	27
GetListOfMeshes	1,1	0,9	0,1
GetListOfAnalyses	58,6	6,9	0,1
BoundingBox	6,4	1,3	0,6
GetListOfMeshes	1,1	0,9	0,4
BoundaryOfAMesh	127,1	22,3	15,1

*Table **6**: Execution times (seconds) for some VQueries of EDM version of the platform run in EDDIE.*

1.3.6 References

[1] I. Foster, Y. Zhao, I. Raicu, and S. Lu. Cloud computing and grid computing 360-degree compared. In Grid Computing Environments Workshop, 2008. GCE â€™08, pages 1 to 10, Nov 2008.

[2] N. Marz and J. Warren, Big Data. Principles and best practices of scalable realtime data systems, Manning Publications Co., Manning Early 2012, September 2013 (est.).

[3] V. Basili, G. Caldiera, and H.D. Rombach, “The Goal Question Metric Paradigm,” Encyclopedia of Software Eng., vol. 2, pp. 528-532, John Wiley & Sons, 1994.

[4] Markus Nick, Klaus-Dieter Althoff, Carsten Tautz, “Facilitating the Practical Evaluation of Organizational Memories Using the Goal-Question-Metric Technique”. KAW’99 – Twelfth Workshop on Knowledge Acquisition, Modeling and Management Track “Evaluation of KE Techniques”

[5] K. Gallagher, A. Hatch, M. Munro, “Software Architecture Visualization: An Evaluation Framework and Its Application”. IEEE Transactions on Software Engineering, vol. 34, no. 2, March/Aapril 2008.

1.4 Potential impact

1.4.1 Strategic value of scalable data analysis and visualization tools within the simulation workflow

The entire simulation world has an impressive impact on any engineering activities, being considered as unavoidable for some industries such as the aeronautic and automotive, where simulation tests are compulsory before creating the first physical prototypes. Some clear examples of that are wind-tunnels, crash tests, structures design, electromagnetic compatibility, ship hydrodynamics, which are areas of expertise of the partners. Any manufacturing industry is a candidate to improve their processes by adopting simulation procedures combined with the most traditional experiments and theory.

The new advances on simulation-based engineering sciences, strongly linked with high performance computing (HPC) and distributed systems, will play an essential role to achieve the grand challenges for engineering, such as new materials, material design, and thus new products and the European innovation.

The current trend in engineering design is to increase physical fidelity (either by considering more scales or linking different disciplines), improve accuracy and robustness, and extend the range of feasible (credible) problem classes. Increased physical fidelity and accuracy (spatial and temporal) in all fields in engineering (civil, mechanical, aerospace, naval, telecom, etc.) naturally imply handling an extremely greater volume of data through HPC facilities, being able to process this growing amount of data in a friendly and understandable way, that is, through intuitive visualization techniques.

Furthermore, high-end computing is increasingly being used to provide extremely large data bases for fast- running engineering models. This area, real-time computing, either interpolates from these detailed data bases, or extracts fundamental modes of the system to obtain a reduced order model, or answers queries, e.g. the FastQuery approach. Consequently, the need for a next generation of faster data transfer computing structures and visualization to achieve these goals is obvious.

Workflow of a CAE engineer

In order to remark the impact of the results in VELaSSCo, a simplified picture of the traditional workflow of a Computer Aided Engineering (CAE) engineer is showed in Figure 29Figure 29: generating the CAD model, simplifying (cleaning the CAD model for simulation purposes, adding, editing, merging and deleting geometrical entities, etc), then assigning properties for the simulation (materials, loads, boundary conditions and other relevant parameters), generating the mesh, running the specific field solver and, finally, getting and visualizing the results.

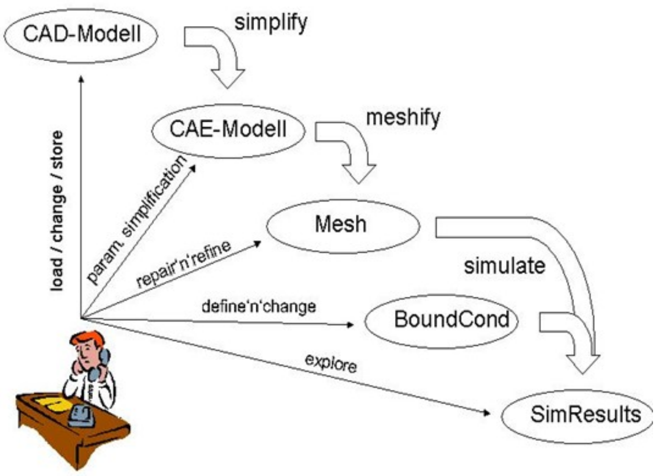

Figure 29. *Simulation tasks for a CAE engineer*

This process can be very iterative depending on the detected flaws at all the levels: some inconsistent geometrical objects, problems in the discretization, and wrong settings on the simulation parameters or undesired results outcome.

Another important aspect to consider is the definition of simulation domain (size) which heavily influences the correctness of the outcome, ability to converge for the solver, etc. The various software tools used are highly dependent on each other. They make use of diverse data models and formats, and can generate huge amounts of data, exchanged in large data files. Besides the consistency issues and error that can appear in such environments, the variety of tools, models and delays incurred raise important challenges in terms of storage, efficiency and usability of the data.

Therefore, sophisticated visualization software is needed that heavily depends on real-time data processing and large-scale storage capabilities.

1.4.2 Impact of VELaSSCo outcomes

The European Digital Agenda mentioned in 2010 that Information and Communications Technologies (ICTs) have driven half of the productivity growth in Europe over the past 15 years, identifying Europe's digital economy as crucial to economic growth and prosperity. For the EU-27, the value added by the ICT industry represented around 5 % of GDP in 2007. At the same time, the ICT sector is the largest R&D investing sector in the EU economy.

The ultimate benefits of research in VELaSSCo can be found in many areas. One can already discern clearly the following benefits of the research proposed. Namely:

- Pushing the state-of-the-art in high-end scientific and industrial computing by enabling the necessary tools for better analysis of its outputs, facilitating thus the work of modellers and easing the way to better simulation models.
- Development of new algorithms and computational techniques for the analysis of large-scale multidisciplinary engineering problems;

More specifically,

- VELaSSCo includes algorithms and tools for advanced querying and analytics applications with interactive response times over petabyte-order data files obtained from simulation of real-scale challenging engineering problems. Given the dimensions of those simulations, data will be distributed by nature over different machines and under different architectures;
- VELaSSCo Platform presents visualization tools; through VELaSSCo, modelling engineers are able to manipulate and interact with the results of their high-performance simulations, both during the run of the simulation (for exploratory analysis of the results and early identification of errors) and for ex-post visualization of the results of the models immediately (or at any other model) the simulation is completed;
- At the production phase, VELaSSCo will contribute to products that are in demand in the Big Data market that will have increased performance at reasonable price and easy accessible;
- VELaSSCo facilitates the necessary tools for European industries to routinely analyses and exploit, via visualization tools, real-scale engineering problems whose outputs increase and computing resources increase;

Academic impact

There is a large need of new tools specifically designed and targeted to be used for the management and analysis of the huge amounts of data resulting from high-resolution real-scale simulations performed on HPC platforms, and more precisely for the upcoming generation of computation infrastructures.

This need is becoming an urgent must-have, since most of today’s available visualization software and platforms have been developed to analyze model outputs run on scalar machines, which often are stored in a single location. Porting all these visualization tools to new platforms able to analyze distributed data run on parallel machines (and therefor, stored in a distributed manner) is not only a matter of scalability. In most of the cases it requires a new reformulation of access, queering and transferring algorithms, which have to be designed to perform at its best under the new infrastructure paradigm.

The following list describes the scientific impact of the algorithms and tools developed in VELaSSCo:

- New efficient, reduced-order post-processing tools including minimization of data transferring and model simplification techniques.
- New visualization and data management platforms with a clear end-user approach to be able to interrogate and visualize huge amounts of distributed data
- New analysis tools for simulation results, together with a higher resolution of the models, allowing a better understanding of the physical phenomena modelled
- New Local Refined Spline technology for representation of scalar and vector fields in general in an interactive way, reducing drastically the stored data.
- Higher quality targeted data through noise-reduction, validation and verification processes in the inject activity, speed of data analysis and data management.

These new developments ease the way towards the peta/exascale of current simulations efforts in the different fields of science and technology (in the framework of the PRACE initiative), easing the route to solving large-scale challenging problems in many different disciplines.

Industrial impact

In the framework of high-end computing, industry has a dual role: firstly, supplying systems, technologies and software services for HPC; and secondly, using HPC to innovate in products, processes and services. Both are important in making Europe more competitive.

Especially for industries and SMEs, access to HPC, modelling, simulation (and its accompanying supporting tools, like the visualization tools developed in VELaSSCo), product prototyping services and consulting is important to remain competitive. The PRACE Action Plan advocates for a dual approach: strengthening both the industrial demand and the supply of HPC.

The actions carried out in VELaSSCo are clearly aligned with the first approach mentioned, and the impacts that the project can have on industries are numerous. In a generic way, the outcome of VELaSSCo is a set of new data managing and visualization tools that allows industries to solve large-scale problem of high economic benefit.

In addition, the project allows a better/faster adoption to the HPC technology in the industry, simplifying the management and visualization of the models. That is, by adopting these new data analytics and visualization tools, most modelling engineers with a very specialized knowledge in HPC will be able to perform large-scale simulations.

Specific industrial impact

The following list enumerates some of the identified specific industrial problems/areas where the outcomes of the project can result in an important leap in their respective fields, positioning European industry in a leading role with respect to international competitors. Additionally, the following applications will also greatly benefit from VELaSSCo results in solving real-scale problems:

- Structural Design and aerodynamic simulation for the aeronautic and automobile industries
- Multiphase/multi-physics simulations, such as fluid-structure interaction in civil engineering

- Particulate/granular systems: Storage, process and transport of granular materials, such as soil and natural/artificial grains for chemical, food, agriculture and mining industry
- Geodynamic/geomechanics simulation
- New techniques for computer-aided optimum design of new engineering materials
- New techniques for computer-aided optimum design of engineering systems considering uncertainties
- New methods for enhanced performance constructions against natural hazards, involving outstanding particulate water flows
- Prediction of noise generated by vehicles (cars, trains, airplanes) and machinery
- Study of the SAR of an object, accounting for all the electromagnetic-thermal couplings
- Computation of the radar-cross-section of a ship or an airplane at realistic Reynolds-numbers
- Study of the hydrodynamic and structural performance of ships, offshore platforms and subsea installations and wind turbine structures in open sea
- Optimum design of selected industrial forming processes: Sheet Metal Forming, Casting, Metal Deposition and Welding
- Life-cycle management of complex and long-lasting products, such as, aircrafts, ships, space equipment and built environment facilities and assets
- Smarter city urban planning using Building Information Model (BIM) technologies
- Long-term data retrieval and archiving for long-lasting engineering products, such as, aircrafts, buildings and offshore structures
- Support the process of obtaining license-to-operate (oil and gas), flight type certificates (aerospace industry) by providing smarter data storage and access facilities adapted to the increasing amount of required simulation data, e.g., for the communication with governmental bodies and authorities
- Basic research concerning Turbulent Flows (e.g. LES calculation around an airplane or car); Chemically Reacting Flows (e.g. LES calculation of mixing and combustion in an engine); Cell and DNA studies, etc.

Initial industrial impact has been measured through a selected User Panel with a membership of scientists and industrial R&D engineers: this ensures both depth and breadth of impact on European stakeholders including SME's, industrial companies and research organisations.

Edinburgh as the Coordinator for the PARDEM ITN (2009-2013) on DEM modelling of granular processes (www.pardem.eu) and T-MAPPP ITN (2014-2018) on multi-scale analysis of multiphase particulate processes leads the exploitation of the outputs of VELaSSCo for the analysis and visualisation of the wide range of industrial and scientific computational modelling conducted in these ITN projects.

In short, the User Panel has been use to actively engage key stakeholders over many industrial and research sectors including pharmaceutics, household products, chemicals, food, equipment manufacture, forestry and construction, aerospace and software houses - it thus provides a most rigorous and effective evaluation of the usability of the resulting software and user interface.

Social impact

The published document “PRACE – The Scientific Case for HPC in Europe2012-2020 from petascale to exascale” indicates that the competitiveness of European science and industry (as mentioned in the previous section) will be jeopardized if sufficiently capable computers are not made available, together **with the associated tools and skilled people necessary to maximize their exploitation.** The set of data managing and visualization tools obtained from VELaSSCo output undoubtedly helps towards the maximization of the exploitation potential of the simulations.

From the social point of view, advances in the field of High-Performance Computing (HPC) will lead us to Enhanced design and manufacturing of economic, safer and environmental friendly industrial products and structures in many engineering sectors: Aero-Space Engineering, Civil Engineering, Naval and Marine Engineering and Surface Transport Vehicles.

Of particular social impact, just to point put a concrete area of engineering that can benefit from the outputs of the project, is the development of new simulation methods able to solve problems of interest for predictive safety of civil constructions to natural hazards. These constructions include: buildings, bridges, harbours, dams, dykes, breakwaters, and similar infrastructure under extraordinary forces due to water hazards such as floods, sea waves and tsunamis, water spills due to the collapse of dams, dykes and reservoirs induced by landslides and earthquakes, among others. The output of this research is essential for enhanced analysis, risk assessment, save-guarding values and performance-based design of constructions, to protect population and infrastructure against natural hazards.

Furthermore, access by researchers, administration and industries to this type on services, and its full exploitation with customized tools such as the ones derived from the results obtained in this VELaSSCo project, can be the difference between life and death; between new jobs and profits or bankruptcy.

HPC modelling and its posterior analysis requires significant expertise to make an efficient utilization of the available resources. VELaSSCo provides room for highly-skilled scientists to develop their work at the highest level of international research in European institutions, while at the same time producing tools to be used by other researchers both in academic research institution and industries. Indeed, the high mobility of researchers suffered in the project, with several of them leaving the project after acquiring some expertise in Big Data for simulation, indicates the high interest of the topic.

Research institutions' most valuable assets are knowledge and know-how, much more than computers, software or mathematical formulations. These assets consist of a wide set of skills that require a rigorous and detailed training. Most of key original background disciplines related to the field of the proposal belong to drastically changing technologies that require a lifelong learning. VELaSSCo will contribute to keeping and fostering that knowledge and know-how of scientists within Europe, thus maintaining the high-quality human resources assets that will ensure European leadership in the Big Data engineering simulation field and giving continuity to the funds and resources dedicated by the EC in this strategic field

1.4.3 Need for a European Approach

The EC document Digital Agenda for Europe (IP/10/581), one of the seven flagship initiatives under the Europe 2020 strategy (IP/10/225), states that "*half of European productivity growth over the past 15 years was already driven by information and communications technologies, and this trend is likely to accelerate*". Under that perspective, a Partnership for Advanced Computing in Europe (PRACE) was launched in 2010 to enable researchers to use super-fast computers located in other countries to make up to 1000 trillion calculations per second for their research projects.

Access to super-fast supercomputers means that research can be conducted incomparably faster, at massive scale and far more accurately. It also allows addressing problems that could not be solved before because they were too large and complex. Therefore the impact of the effort is huge. The final goal is to enable high impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society.

The global push to HPC and closer technologies (Cloud Computing and exascale, to mention two of the most important ones), is at its highest, and in a fierce race to win, Governments are competing to deliver the needed funding to build these systems. Just recently the European Commission (COM(2012)45, High- Performance Computing: Europe's place in a Global Race), announced it is doubling its investment in the push for exascale computing from 630 M€ to 1200 M€. This announcement, in a context of general budgetary contention imposed by the current crisis scenario, proves the importance that Europe is given to this field, in the certainty that High Performance Computing is a crucial enabler for European industry and for more jobs in Europe. As an EC's responsible for the Digital Agenda said "We've got to invest smartly in this field because we cannot afford to leave it to our competitors".

Under this strategic framework, VELaSSCo outcomes contribute with the full exploitation of the existing Big Data, HPC and Cloud Computing infrastructures. Therefore, VELaSSCo consolidates the EU's strong position in HPC related technologies by ensuring the availability of high quality visualization tools for engineering simulation users.

1.5 Contact details.

Web: www.velassco.eu
Twitter: *@VELaSSCoProject*
LinkedIn: www.linkedin.com/in/velassco/
Youtube: *VELaSSCo EU Project* channel
Scientific Coordinator
Dr. Abel Coll
CIMNE (International Center of Numerical Methods in Engineering)
Email: projectes@cimne.upc.edu