Scientific Report

First name / Family name	David Mera Pérez
Nationality	Spain
Name of the *Host Organisation*	Masaryk University Faculty of Informatics and Institute of Computer Science
First Name / family name of the *Scientific Coordinator*	Pavel Zezula
Period of the fellowship	01/10/2013 to 31/03/2015

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

According to my original research program, the main goal of my ERCIM project was the study and development of a generic and distributed system for processing large multimedia datasets. Specifically, the developed system should process heterogeneous data from heterogeneous data sources. Moreover, it should take advantage of the available resources and adapt its parallel infrastructure during runtime according to the needs of the running tasks.

Recent hardware and software developments allow us to produce unprecedented amount of data, which used to be poorly filtered before being stored. Consequently, the amount of unstructured datasets grows exponentially every year and the gap between the available data and the data that organizations can process is getting bigger. This Big Data phenomenon is critical from the multimedia point of view, where the RAW data must be processed in order to uncover useful information. The necessary processing tasks are highly time consuming and they should be parallelized for dealing with large datasets. The development of distributed approaches demands high computer skills. Regularly, organizations develop their 'ad-hoc' approaches focused on specific scenarios and problems. Typically, these approaches have poor reusability.

My ERCIM project was composed of four main activities:

Review of the state of the art

As was planned in the research program, the first months were focused on the review of state of the art of the Big Data phenomenon as well as on the multimedia data processing, with special attention to the stream processing.

Empirical evaluation of different approaches for processing large datasets

There are several approaches which can be used to process huge multimedia collections. The use of one instead of the others should be carefully considered according to the processing scenario in order to get the best performance. During this phase, I have been running a deep empirical evaluation of several distributed processing approaches based on Grid Computing, Apache Hadoop and Apache Storm. These approaches were checked in different scenarios in order to know which one was the most appropriate in each case.

Development of a distributed system for processing large multimedia datasets

The main part of the fellowship was focused on the development of a distributed and adaptible processing system based on the Apache Storm framework. The developed system does not demand specific skills about distributed computing. It can be deployed as a service for automatically processing user jobs. The system is able to create a distributed job from the basis of a sequential user job and adapts its parallelization level according to the available resources. The system development was composed by three main phases:

- *Input job management system*: a job definition language was created to specify user jobs. A user job definition is composed by both the data source and the sequence of algorithms that must be applied over the target data. The input job management module gets user jobs and creates distributed processing graphs. Moreover, it also creates streams of data from the specified data source and feeds the graphs with them.
- *Monitoring system*: a system module monitors the distributed job and modifies its parallelization level according to the data stream as well as the available resources.
- *System deployment*: the first prototype was deployed over few resources, which

were part of the research group. However, the last system version has been deployed over a virtual cluster. Cluster nodes were virtualized through the Infrastructure-as-a-Service cloud provided by the MetaCentrum Virtual Organization. This infrastructure allows the use of the system in a real environment.

Dissemination of results

According to the research program, it was expected one conference paper and one journal paper in a conservative approach and two journal papers in desirable perspective. At this moment, a conference paper was published and one journal paper was already submitted. A second journal paper is almost ready and we expect to submit it the following month.

The dissemination of results was also carried out through several meetings during the Research Exchange Programme as well as in several internal DISA seminars.

This fellowship allows me to go deep in a research line, which I plan to follow in the next years. Currently, I stay in contact with Professor Pavel Zezula and his research team and we aim to cooperate and work together in the future. We also plan to work in the development of Marie Sklodowska-Curie Innovative Trainning Network through the links created during the ERCIM fellowship as well as my previous ones.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Published papers

- D. Mera, M. Batko, P. Zezula, Towards fast multimedia feature extraction: Hadoop or storm, *in: Proceedings of the 2014 IEEE International Symposium on Multimedia, IEEE Computer Society*, 2014, pp. 106–109.
  Abstract: The current explosion of data accelerated evolution of various content-based indexing techniques that allow to efficiently search in multimedia data such as images. However, indexable features must be first extracted from the raw images before the indexing. This necessary step can be very time consuming for large datasets thus parallelization is desirable to speed the process up. In this paper, we experimentally compare two approaches to distribute the task among multiple machines: the Apache Hadoop and the Apache Storm projects.

Submitted papers

- *D. Mera, M. Batko, P. Zezula***,** Speeding up the multimedia feature extraction: the Big Data approach, *Multimedia Tools and Applications*, 2015.
  Abstract: The current explosion of multimedia datasets is significantly increasing the amount of potential knowledge. However, to get to the actual information requires to apply novel content-based techniques which it turn require time consuming extraction of indexable features from the raw data. In order to deal with large datasets, this task needs to be parallelized, however, there are multiple approaches to choose from, each with its own benefits and drawbacks. There are also several parameters that must be taken into consideration, for example the amount of available resources, the size of the data and their availability. In this paper, we empirically evaluate approaches based on Apache Hadoop, Apache Storm, and Grid computing employed to distribute the extraction task over an outsourced and distributed infrastructure

Papers to submit

Title: *D. Mera, M. Batko, P. Zezula*, Towards an automatic and adaptible distributed system for processing stream data
Abstract: The current explosion of data has given rise to existence of the Big Data

phenomenom, where many organizations are not able to process or analyze available datasets. There are several frameworks which allow distributing the processing of large datasets across clusters of computers. However, they require a development learning period and they also demand a careful configuration tuning in order to improve their performance. Generally, organizations develop 'ad-hoc' applications with poor reusability for dealing with specific scenarios. This paper presents a distributed and adaptible stream data processing system based on the Apache Storm framework. The system does not require distributed computing skills, since it is able to create distributed topologies from the basis of user sequential jobs. Moreover, a monitor system checks the status of the running topology during runtime, in order to modify its level of parallelism according to the rate of stream data and the available resources. The system has been tested with applications focused on processing streams of multimedia data.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

- ERCIM Annual Seminar, Athens (Greece). October 31-November 1, 2013
- IEEE International Symposium on Multimedia (ISM2014), Taichung (Taiwan). December 10-12, 2014.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

- **Research center:** Information Science and Technology Institute (ISTI) of National Research Council of Italy (CNR). Pisa (Italy).
  Local scientific coordinator: Fausto Rabitti
  Duration: 1 week from 17th to 21st of March 2014
  Visit summary: There are several groups working in the ISTI, some of them with research lines closely related to my project and others with lines farer. During that week, I presented my ERCIM project in a seminar, where I got several useful insights after a useful discussion. Moreover, I set several individual meetings with the different ISTI research groups in order to know in detail their research lines as well as to discuss my research project from their viewpoints.

- **Research center**: VIPER: Multimedia Information Retrieval, Computer Science Center. University of Geneva. Geneva (Switzerland)
  Local scientific coordinator: Stephane Marchand-Maillet
  Duration: 1 week from 16th to 23st of November 2014
  Visit summary: I had the opportunity to attend to one of the seminars of the research team, where they showed me some of their research lines. Moreover, I had several individual meetings with some of the researchers, where I could show in detail my research work as well as discuss its applicability in their research areas. During that visit, I also had the opportunity to set meetings with other two research groups in Switzerland. The first meeting was with the Distributed Information System Laboratory (LSIR) in the University of Lausanne. The second one was with the eXascale InfoLab research group of the University of Fribourg. Both groups work in research areas close to my ERCIM project. I presented my research work and the group researchers provided me useful insights. Moreover, I could also know some of their research projects.