BigFoot: Big Data Analytics of Digital Footprints

<table>
<thead>
<tr>
<th>Project name</th>
<th>BigFoot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project ID</td>
<td>FP7-ICT-ICT-2011.1.2 Call 8 Project No. 317858</td>
</tr>
<tr>
<td>Working Package Number</td>
<td>WP6</td>
</tr>
<tr>
<td>Deliverable Number</td>
<td>D.6.2.3</td>
</tr>
<tr>
<td>Document title</td>
<td>Dissemination and Promotion Activities Report (Finale version)</td>
</tr>
<tr>
<td>Document version</td>
<td>3.0</td>
</tr>
<tr>
<td>Author</td>
<td>GRIDP, ALL</td>
</tr>
<tr>
<td>Date</td>
<td>25/09/2015</td>
</tr>
<tr>
<td>Status</td>
<td>Public</td>
</tr>
</tbody>
</table>
## Revision History

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Description</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>25/09/2015</td>
<td>3.0</td>
<td>Document update</td>
<td>F. Gluszak</td>
</tr>
<tr>
<td>28/09/2015</td>
<td>3.0</td>
<td>Document review</td>
<td>P. Michiardi</td>
</tr>
</tbody>
</table>
Executive Summary

This document is a report summarizing dissemination, communication, collaboration and open-source activities conducted by the consortium during the first, the second and the final, third, year of the BIGFOOT project.

<table>
<thead>
<tr>
<th>Activity Type</th>
<th>Y1 Count</th>
<th>Y2 Count</th>
<th>Y3 Count</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliverables</td>
<td>15</td>
<td>10</td>
<td>8</td>
<td>33</td>
</tr>
<tr>
<td>Seminar, Tutorial, Talks</td>
<td>19</td>
<td>11</td>
<td>11</td>
<td>41</td>
</tr>
<tr>
<td>Scientific publications</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td>39</td>
</tr>
<tr>
<td>Press Release</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Demos</td>
<td>2</td>
<td>2</td>
<td>11</td>
<td>15</td>
</tr>
</tbody>
</table>
Contents

1 Introduction 5

2 Scientific Publications 6
   2.1 Third year of the project 6
   2.2 Second year of the project 8
   2.3 First year of the project 9

3 Communications 12
   3.1 Seminars, Talks, Tutorials and Demonstrations 13
      3.1.1 Year 3 13
      3.1.2 Year 2 15
      3.1.3 Year 1 16
   3.2 BigFoot Video Project 17
      3.2.1 Script Video 1: “BigFoot Overview” 18
      3.2.2 Script Video 2: “Technical Highlights” 19
      3.2.3 Script Video 3: “BigFoot Applications” 20
      3.2.4 Further details on video production 20

4 Collaborations 22

5 Open Source Software 30

6 Conclusion 36
1 Introduction

Clearly, the dissemination of results achieved in BIGFOOT is of extreme importance. As stated in the DoW, partners committed to present their fundamental and experimental results at key scientific conferences and journals.

In addition, this document reports the efforts that have been made in the first, second and third year of the project to communicate on project results, to establish collaborations with other relevant projects and their fruits, and the on-going effort to publish software results as open or free source projects.

This is the final release of a live deliverable, that has been updated and enhanced at each project milestone, during three years of the BIGFOOT project. For clarity, all larger extensions include indications of the project period they apply to.
2 Scientific Publications

This Section contains a list of publications for the first and second year of the project, including conference, workshop and journal papers, and book chapters.

2.1 Third year of the project.

- Shen, Yun; Thonnard, Olivier; Vervier, Pierre-Antoine; Dacier, Marc. *Scalable Multi-Criteria Data Clustering for Big Data Security Intelligence Analytics*. IEEE Transactions on Big Data 2015, revision under review.

  Relevant work-packages WP2

- Han, Yufei; Shen, Yun. *Positive unlabelled feature selection using constrained affinity graph embedding*. CMStatistics 2015, International conference of the ERCIM WG on Computational and Methodological Statistics, accepted and will appear soon in the proceedings.

  Relevant work-packages WP2

- Sarlis, Dimitrios; Papailiou, Nikolaos; Konstantinou, Ioannis; Smaragdakis, Gerogios; Koziris, Nectarios. *Datix: A System for Scalable Network Analytics*. CCR 2015, Technical review track

  Relevant work-packages WP3 and WP4


  Relevant work-packages WP4


  Relevant work-packages WP3

- R. Appuswamy, M. Olma and A. Ailamaki *Scaling the Memory Power Wall with DRAM-Aware Data Management*. DaMoN’15, Proceedings of the 11th International Workshop on Data Management on New Hardware.
Relevant work-packages WP4


Relevant work-packages WP4


Relevant work-packages WP3

- Phan, Duy-Hung; Hoang-Xuan, Quang-Nhat; Dell’Amico, Matteo; Michiardi, Pietro. *Efficient and self-balanced ROLLUP aggregates for large-scale data summarization*. **4th International IEEE Congress on Big Data**, June 27-July 2, 2015, New York, USA.

Relevant work-packages WP3


**Collaboration** GRID and EUR, Relevant work-packages WP2, WP3


**Collaboration** RMA, SYM and EUR, Relevant work-packages WP2, WP3


Relevant work-packages: WP2
2.2 Second year of the project


Relevant work-packages WP4


**Collaboration** RMA, SYM and EUR, Relevant work-packages WP2, WP3


**Collaboration** EUR and EPFL, Relevant work-packages WP3, WP4


Relevant work-packages: WP3


Relevant work-packages: WP3


Collaboration SYM and EUR, Relevant work-packages WP2, WP3


Relevant work-packages WP2, WP3

FP7-ICT-ICT-2011.1.2 Call 8 Project No. 317858
  Relevant work-packages WP4
  Relevant work-packages WP3
  Relevant work-packages WP4
  Relevant work-packages WP5
  Relevant work-packages WP5
  Relevant work-packages WP5

### 2.3 First year of the project.

  Relevant work-packages: WP2, WP5
Relevant work-packages: WP5


Relevant work-packages: WP5


Relevant work-packages: WP5


Relevant work-packages: WP5


Relevant work-packages: WP5


Relevant work-packages: WP3

- Matteo Dell’Amico. *A Simulator for Data-Intensive Job Scheduling*.

Relevant work-packages: WP2, WP3


Relevant work-packages: WP5

Relevant work-packages: WP5


Relevant work-packages: WP5


Relevant work-packages: WP5


Relevant work-packages: WP3, WP4

• Mario Pastorelli, Antonio Barbuzzi, Damiano Carra, Matteo Dell’Amico, Pietro Michiardi. *Practical Size-based Scheduling for MapReduce Workloads.*

Relevant work-packages: WP3

Others:

• Georgios Smaragdakis (TU Berlin) is one of the contributing experts for the EU Roadmap for Advanced Cloud Technologies under H2020.
3 Communications

To ensure the visibility of BIGFOOT, a website (http://bigfootproject.eu) has been set up as of the early stages of the project. This website gives a clear description of BIGFOOT objectives, of the academic and industrial partners and of the two applications areas driving the project, namely Cyber-Security and Smart Grids. Additionally the website has a public section for Deliverables and consolidates the publication list of project partners, and also links to relevant open-source projects. Also visible on the front page of the website is the Twitter time-line of the BIGFOOT Twitter account (@BigFoot_project).

The project website has been kept up to date throughout the entire duration of the BIGFOOT project, and it has been updated several times during the third year.

The new website features added during the third year include several video recordings of talks and presentation done by members of the consortium that has been made visible directly within the BIGFOOT website pages (see Figure 1), and the new sub-page dedicated to downloading and installation of the entire BIGFOOT solution (see the chapter 5. Open Source Software for more details).

Finally, this Section covers the major effort done in the Third year of the project to maximize project visibility through a series of 3 professional videos.

Figure 1: Video presentations available on bigfootproject.eu website
3.1 Seminars, Talks, Tutorials and Demonstrations.

In what follows, we present the list of events where BIGFOOT was presented either in the form of talks and seminars or in a more practical form such as demos and tutorials.

3.1.1 Year 3

- **Talk** *Universal Exposition Milan 2015*. Planned participation of GridPocket (as a winner H2020 SME Innovation Instrument support) to the side event of Expo Milano, involving talks and demonstration of smart grid data processing to international companies, October 2015.

- **Demo** *Cop21 Lyon*. Demonstrations and talks by GridPocket at the Lyon EDF pre-event of the Global Climate Summit organized by United Nations, Lyon (France), July 2015.

- **Demo** *Innovative City*. Demonstration of GridPocket’s smart grid data processing solution at the international conference and trade show dedicated to smart cities in Nice (France), June 2015.

- **Talk** *MedCop21*. Presentation on smart grid data processing technologies to professionals, politics and institutions of the Mediterranean region countries in Marseille, June 2015.

- **Demo** *SmartGrid Paris*. GridPocket technology demonstration, including data processing platform at SmartGrid Paris professional tradeshow, May 2015.

- **Talk** *Cutting Edge*. Bigfoot presentation at Symantec conference “Cutting Edge” in Tallinn (Estonia), May 2015.

- **Talk** *WITX Ignite*. Introduction to BigFoot and other research projects by GridPocket at the WITX (Informations Wilmington Technology Exchange) conference at UNCW (University of North Carolina Wilmington, USA), April 2015.

- **Talk** *RTCC*. Presentation on European clean tech industry, and research projects, including BigFoot at the workshop with Research Triangle CleanTech Cluster at the RTC park in NC, USA, April 2015.
• **Talk** *SKEMA Raleigh*. Presentation on Smart Grid technologies, and related research projects, including BigFoot to students and staff of SKEMA business school in Raleigh, NC, USA), April 2015.

• **Talk** *ASPRoM*. Presentation on BigData, including BigFoot project at the professional conference “Smart networks for electricity, gas and water” organized by ASPROM association, Mars 2015.


• **Demo** *BigFoot AaaS remote demo for FastWeb*. Pietro Michiardi, April 2015.

• **Demo** *BigFoot AaaS demo for ST Microelectronics*. Pietro Michiardi, March 2015.

• **Demo** *BigFoot AaaS demo for Eurecom Industrial Advisory Board*. Pietro Michiardi, February 2015.

• **Demo** *NoDB: Efficient query processing on raw data files*. EPFL, Ecoaloud, 2014.

• **Demo** *NoDB/ViDa: accès efficace aux données brutes*. EPFL, Open House Day at the IC School (22/11/2014).

• **Talk** *BigFoot AaaS for OpenWorld Forum*. Matteo Dell’Amico, November 2014

• **Demo** *OneM2M workshop*. Demonstration of GridPocket energy data management platform, including BigFoot results, at the OneM2M standard global interoperability event in Mandelieu (France) in partnership with international technology partners including Vodafone, Qualcomm, SierraWireless, December 2014.

• **Seminar** *GreenIT*. Presentation of BigFoot project along other tools for energy efficiency to graduate student of Eurecom, December 2014.

• **Talk** *DigiPolis*. Introduction to Big Data security issues in relation to Smart Grid, including brief presentation of BigFoot project at professional conference DigiPolis in Montbeliard (France), November 2014.

• **Talk** *European Utility Week Conference*. Talk about GridPocket’s products including TSL - Time Series Labs at the conference during European Utility Week in Amsterdam, November 2014.
• **Demo** *European Utility Week*. Demonstration of GridPocket’s products including TSL - Time Series Labs at the annual smart grid event European Utility Week in Amsterdam, November 2014.

### 3.1.2 Year 2

- **Demo** *Reactive and Proactive Sharing Across Concurrent Analytical Queries*. Iraklis Psaroudakis, Manos Athanassoulis, Matthaios-Alexandros Olma, Anastasia Ailamaki. *ACM SIGMOD 2014*

- **Talk** *Size-Based Scheduling: From Theory To Practice, And Back*. Matteo Dell’Amico, SophiaTech networks seminar, Sophia-Antipolis, France, April 24th 2014

- **Seminar** *CloudWatch Concertation Meeting 2014*. BigFoot presentation, and SME GridPocket presentation, September 2014 (GRID-POCKET)

- **Talk** *European Patent Office*. Bigfoot presentation, June 2014 (EUR)

- **Seminar** *HPCloud*. Bigfoot presentation, May 2014 (EUR)

- **Seminar** *SophiaTech Networks Seminar*. BigFoot presentation, April 2014 (EUR)

- **Talk** *OCEAN Collaboration Meeting*. BigFoot presentation, April 2014 (EUR)

- **Seminar** *EDBT 2014*. BigFoot Presentation in the Special Track on EU projects, April 2014 (EUR)

- **Seminar** *FUI 2014*. Bigfoot presentation and poster, April 2014 (EUR)

- **Seminar** *Concertation Meeting 2014*. Bigfoot presentation, March 2014 (EUR)

- **Seminar** *Symantec Cutting Edge Conference series 2014*. Lessons Learned from Clustering 50 Million Spam Emails using MapReduce., Dublin/Tallinn/Mountain View/Culver City, Feb-March 2014 (SYM, Yun Shen)
- **Seminar** *Symantec Cutting Edge Conference series 2014.* Visual Big Data Analytics., Dublin/Tallinn/Mountain View/Culver City, Feb-March 2014 (SYM, Olivier Thonnard)

- **Talk** *ESSEC Paris lecture.* June 2014, (GRIDPOCKET)

- **Talk** *Telecom Bretagne lecture.* February 2014, (GRIDPOCKET)

- **Talk** *Demos at Smart Grids Exhibition in Paris 2014.* June 2014, (GRIDPOCKET)

- **Talk** *Tradeshow in Innovative City Convention 2014.* June 2014, (GRIDPOCKET)

### 3.1.3 Year 1


- **Seminar** *Big Data workshop.* Anja Feldmann (TUB) takes part in this workshop featuring other speakers such as Volker Markl (TUB), John Hopcroft (Cornell). Anja’s talk covers IXP big data analysis. Berlin, 29 September 2013


- **Talk** *Le Big Data avance à grands pas: un exemple dans les Smart Grids (Big Data makes great strides: an example in Smart Grids).* ASPROM “Le potentiel et les défis du Big Data” (potentials and challenges of Big Data) conference, UIMM Paris, 2-3 July 2013


- **Seminar** *The BigFoot Project.* Project presentation and research directions, Orange Labs, May 2013

- **Seminar** *The BigFoot Project.* Project presentation and research directions, AMADEUS, May 2013
3.2 BigFoot Video Project

During the third year, a series of 3 short (about 1min 30 sec each) animated cartoon-type videos has been produced. The objective of these videos is
make promotion of BigFoot technologies among potential users and integrators. They will also serve to convey the vision of BigFoot and educate about possible use cases of big data technologies. Additionally, a 4th video has been produced, that includes direct interview of some of the BIGFOOT project members.

In order to achieve professional result, the BIGFOOT team has hired an experienced video and animation production company (Mediafaune). This company has been selected between half a dozen pre-qualified candidates based on the quality of its offering. During the first phase of the project, the consortium has agreed on the overall objectives and storyboard, that splits the content into three movies:

- BigFoot overview - an introduction to project vision and objectives, addressing large, non-specialized, audience
- Technical highlights - quick summary of major technical and scientific innovations, addressing IT engineers and researchers
- Applications - presentation of two practical uses cases in security and energy, addressing audiences interested in practical usages of big data technologies, as well as non-IT specialists

Following these pre-defined objectives, the work has been focused on design of the graphical elements, and copywriting of the script, that will be read by a professional speaker. In order to help viewers build a personal relation to the video’s subject (BIGFOOT), a main character has been introduced. This character, named Eva, is a passionate data scientist. She believes big data solutions are important tools that enable positive evolution of our lives, just as other technological and cultural inventions did in our history, in particular during the dynamic renaissance period. The main challenge for Eva is to merge the growing demand for data analytics with lack of elasticity in the existing corporate IT systems. This is the reason why she engaged with the BigFoot project, and wish to share her vision with other people.

The final version of the script consists of the following three sections.

3.2.1 Script Video 1: “BigFoot Overview”

Hello, my name is Eva and I am a data scientist. I use Big Data to solve problems in physics, economics, environmental conservation and many other fields. Since the Renaissance times, my predecessors bravely scrutinized
books and precision instruments in order to understand and improve their world. Now, we live in an age of massive digital information and unprecedented global challenges. To cope with them, we have put in place Big Data software tools, which can discover patterns, predict trends and learn from large data sets. Let me present you our latest tool, Project BigFoot. With the need for BigData solutions, IT departments have more and more difficulties to handle the growing number of computing and analytics demands from all individual data scientists. Businesses keep on adding infrastructure and manpower to solve the increasing number of analytics problems. Unfortunately, the resulting technical and organizational complexity impedes the value that data science could bring. Our open source project, BigFoot, enables easy consolidation of big data within existing corporate information systems. Thanks to BigFoot, data scientists can enjoy immediate access to shared private cloud resources, without a need to wait and make new investments. We hope this “analytics as a service” approach will enable novel ideas to evolve even faster! Ready to join us? Help us build the community.

3.2.2 Script Video 2: “Technical Highlights”

The goal of BigFoot is to provide Analytics as a Service on the private cloud. This flexibility is achieved thanks to virtualization of processing, storage and networks. We enable automatic deployment and consolidation of big data applications with private clouds along with support of high level coding languages. This virtual Analytics as a Service infrastructure can be automatically deployed and self tuned with integration in the OpenStack environment. All you have to do is focus on your data analytics goals. BigFoot helps you work efficiently and interactively by scheduling queries and managing tasks according to their processing time and priority. In addition, when queries are similar, BigFoot helps you save on resources by packing them together and letting them share work. BigFoot allows data analysts to interactively explore and access raw data without having to move them from the private cloud to any external database. This is made possible thanks to the in-situ results indexation ensured by our innovative DiNoDB [read : dee-no-deebee] data base system. RAW-labs, start-up company, provides DiNoDB users with a global business-grade support. And this is what Im here for! - Faster time-to-insight and lots of EUREKA moments for data scientists !! Ready to join us? Help us build the community.
3.2.3 Script Video 3: “BigFoot Applications”

I’d like to present to you 2 real life applications of BigFoot technology in the area of energy and security. When designing energy efficiency systems, the challenge is to quickly analyze large data sets coming from millions of smart grid sensors. GridPocket Energy analytics applications, running on top of BigFoot, can efficiently predict energy consumption, manage low carbon production, disaggregate loads, and help users to reduce their energy bills. BigFoots highly scalable and interactive data analytics enable energy utilities across Europe to take global leadership in the deployment of large scale critical smart grid applications. The massive amount of data that security companies collect makes it increasingly difficult for security analysts to correlate and attribute internet threats. To detect cyber attacks, Symantec requires advanced analytics tools like BigFoot to perform multi dimensional clustering analytics. That allows it to find correlations between apparently unrelated events. Bigfoot enables more scalable data mining and data fusion algorithms for detection and analysis of attack activity. If you think BigFoot can help you in your research and business applications, call us, join us, contribute! Help us build the community!

3.2.4 Further details on video production

The animated video features introduction scenes, where Eva explains the vision and objectives of the project standing in a library-inspired decorations (see Figure 2).

The introductions are followed by detailed technical animations, showing various aspects of data science and BIGFOOT technologies (see Figure 3). There are about 10 technical animations for various BIGFOOT features and usages in the overall series.

Each movie is finished by a “call for action”, eg. contacting BIGFOOT consortium, checking the website or downloading open source software. The video contains also logos of the co-financing institution (European Commission), project partners and code repositories, where BIGFOOT software can be found (see Figure 4).
Figure 2: Early pre-production shot from the BigFoot video - Library scenery

Figure 3: Shot of video explaining data scientist work
4 Collaborations

In the first year, the consortium has been active in establishing collaborations with other relevant projects. In particular, this Section reports on two activities that resulted in highly-visible and successful dissemination activities, which were achieved throughout the cooperation with the FP7-IP UNIFY and the FP7-IP mPlane projects.

In the second year, the consortium has established multiple collaborations, and participated to Concentration Meetings organized by the European Commission in June and September 2014. Current collaborations are described in more detail in this Section.

In the third year, the consortium has consolidated the multiple open collaboration to arrive at concrete contributions by external partners, and to increase project visibility, for example by using it as a basis of collaboration with new H2020 projects, and for demonstration of related use-cases. Such collaborations were established both with existing and new members, as detailed in the following Sections.
YEAR 1: Workshop on Distributed Cloud Computing (DCC 2013)

The International Workshop on Distributed Cloud Computing (DCC) is interdisciplinary and touches both distributed systems as well as networking and cloud computing. It is intended as a forum where people with different backgrounds can learn from their respective fields and expertise. The workshop attracts both industry-relevant papers as well as papers from academic researchers working on the foundations of the distributed cloud. The front-page, with important dates and news, is illustrated in Figure 5.

![Figure 5: DCC 2013, Workshop front-page.](image)

The workshop features one keynote speech by Rick McGeer, Distinguished Technologist at HP Enterprise Services, Palo Alto, USA, with the title: “Software-Defined Network and Distributed Clouds: Abstractions for the Network Control Plane”.

- WebSite: [https://sites.google.com/site/dcc2013workshop/](https://sites.google.com/site/dcc2013workshop/)
- Dates: December, 2013
• Number of accepted papers: 10

YEAR 1: First Ph.D. School on Big Data and Applications

The goal of this school is to provide a comprehensive view on recent topics and trends in distributed systems and cloud computing. It covers the software techniques employed to construct and program reliable, highly-scalable systems, with a particular focus on data-intensive computing systems.

Specifically, the school covers the MapReduce programming model, its connection to relational algebra, and high-level programming models that build on MapReduce; in addition, the course delves into the details of the underlying execution framework that supports and execute parallel MapReduce programs, including distributed file-systems and the Hadoop implementation. The course is complemented by a series of practical, hands-on exercises executed on a small-scale cluster and in individual virtual machines, in which students learn the tools to program in MapReduce and Pig; such exercises are drawn from real-world case studies, including for example network-traffic analysis. The front-page of the school Web Site is reported in Figure 6.

Since the school took place before the end of the first year of the project, it is possible to draw some statistics about its impact. The initial enrollment threshold of 15-20 students had to be revised due to the success of the school, which counted more than 60 participants, including Ph.D. student, Engineers from the industry, and Professors from the institutions that co-organized the school.

• WebSite: http://www.ict-mplane.eu/public/phd-school-mapreduce
• Dates: September, 2013
• Number of participants: 60+
• Number of countries involved: 7+

YEAR 1, 2 AND 3: ORANGE / ORANGE LABS Collaboration

In the first year of the project, ORANGE LABS was accepted as a member of the IAB of BigFoot: the related activity, covered mainly presentations about the BigFoot project vision, at all levels of the BigFoot software stack. Several meetings were set up to understand common interest on the research and development axes defined in the BigFoot project: at the end of the first
year, a common research direction, within the scope of WP5, was defined. The main topic of the collaboration was defined to tackle performance issues due to system and network virtualization, measurement tools to fingerprint private cloud deployments based on Apache OpenStack, and the Analytics-as-a-Service layer of BigFoot.

Concretely, the collaboration really started in the first half of second year of the BigFoot project: a Ph.D. student was hired and funded by an industrial scholarship to work on the topics outlined above. Initially, the Ph.D. student (Son-Hai Ha) worked at Orange and Orange Labs premises in Paris and Sophia-Antipolis respectively, to test, validate and use “in production” one of the software tools developed within WP5, namely OsMeF. The current plan is to extend OsMeF with the ability to perform disk I/O measurements, in addition to network profiling. Additionally, the Ph.D. student focused on a validation of the service layer of BigFoot, by deploying the Sahara plug-in developed in WP5 in ORANGE LABS, and by studying the workloads generated internally. The Ph.D. student focuses on the topic of service optimization (which is covered in the third year of the BigFoot project): the goal is to offer automatic and self-tuned services by, in
addition to what is done in BigFoot, considering the impact of large-scale network topologies that are typical of telecommunication operators such as ORANGE.

In the last year of the project, the Ph.D. student focused on one important question that was also discussed during the second annual review meeting of the BigFoot project: what are the implications of virtualization? What are the overheads? To this end, the joint work between the BigFoot consortium (and especially partners involved in WP5) and ORANGE LABS, materialized in a large scale measurement campaign that has been reported in D.2.4.3, and that is currently under submission in a top-tier conference.

**YEAR 1 AND 2: Royal Military Academy**

Starting from the end of the first year of the BigFoot project, a Capitain (Thibault Debatty) affiliated with the Royal Military Academy (RMA) in Belgium contacted Symantec to work on a project related to SPAM classification. Since then, Thibault has planned to enroll as a funded Ph.D. student in a co-supervision between Eurecom and RMA: the official enrollment date is set for the end of October 2014. In the meanwhile, Thibault collaborated with Symantec and Eurecom to work on additional components of the Scalable Machine Learning Library, with the ultimate goal of gaining a better understanding of the algorithmic problems stemming from the analysis of large amounts of data. Ultimately, the topic of the Ph.D. thesis of Thibault will be on the design of new resource allocation and scheduling mechanisms for distributed data processing systems that execute complex workloads generated by Machine Learning algorithms.

We summarize here the contributions to BigFoot from the RMA collaboration:


The collaboration has continued also in the third year, albeit on topics that are not covered by the BigFoot DOW. In synthesis, the Ph.D. subject has been better defined, and will cover approximation algorithms to build scalable clustering algorithms. Such research can be beneficial for the industrial members of the BigFoot consortium, and several meetings to present results have been organized, especially with Symantec. Nevertheless, we don’t report results in WP2 nor in WP6, to mark the difference that exists between the various lines of research.

YEAR 1 AND 2: AMADEUS

The collaboration with Amadeus started in the first year of the project, since when they joined the Industrial Advisory Board. Initially, the internal contact was interested in the infrastructure side of BigFoot. With time, the interest from Amadeus shifted to the application level (and consequently, the contact point has changed as well, as reported in this Deliverable), considering the BigFoot software stack as a perfect fit for their internal research agenda.

Throughout a series of face-to-face meeting, Eurecom demonstrated the BigFoot platform and the TreeLib library that focus on decision trees. Specifically, Eurecom demonstrated the versatility of TreeLib by addressing a long-lasting problem in Amadeus, that is related to the prediction of passenger flow in some specific zones of an airport. Eurecom showed that passenger flow is highly related to aircraft delays, and showed (using publicly available data), how to predict such delays with TreeLib.

Today, the collaboration has been formalized: a Master Thesis has been defined (and started in July 2014) on the topic of revenue assurance. Essentially, the goal of the Thesis is to mine hotel reviews available through the Amadeus global network (e.g. Amadeus closely work with Booking.com and many other companies) and forecast customers’ rating and their evolution with respect to changes in the amenities offered by hotels. The Thesis is currently on-going, and will finish by the end of the year 2014.

In the third year of the BigFoot project, the collaboration with AMADEUS has intensified. Nevertheless, we chose not to report progress in WP6 as the terms of the collaboration go beyond the project end. Precisely, a bilateral contract with AMADEUS has been established, with focus on Anomaly and Fraud Detection techniques. Such contract will start in January 2016.
YEAR 1 AND 2: SWISSCOM

The collaboration with SWISSCOM started as a Master Thesis project from March to August 2014. The project consisted in solving the problem of “customer location inference”, whereby the goal is to infer the next location of a mobile user for Quality of Service purposes. In practice, this work used the TreeLib machine learning library developed in WP2 (with collaborations in WP3) to perform the forecasting task. The results obtained in this project were encouraging, with a prediction accuracy close to 90%.

Currently, Eurecom is in discussion with SWISSCOM to continue the collaboration, by creating a private tenant on the BigFoot platform at Eurecom to expose the BigFoot software stack to research engineers at SWISSCOM. The work will focus on the topic of the design of scalable machine learning algorithms.

YEAR 2: SeCludIT

The collaboration with SeCludIT, a start-up company in the ICT security business, started simply because the company needed the Analytics-as-a-Service features of the BigFoot software stack. As such, initially, security researcher of the company used (and continue to use, at the time of writing) the BigFoot software stack to perform research in the domain of secure data de-duplication.

Currently, Eurecom and SeCludIT are defining a joint research agenda to use the BigFoot software stack (and the TreeLib library) to predict the risk index of virtual machines in the public cloud, such as Amazon Web Services.

YEAR 3: mPlane Project

As a continuation of the collaboration activities between BigFoot and mPlane, the BigFoot platform is going to be used in the final review meeting of the mPlane project to support demonstration activities in the domain of large-scale Quality of Experience monitoring.

The demonstrator focus is on algorithmic techniques to discern and pinpoint root causes of poor Web browsing performance. This is achieved by collecting a vast amount of data using software probes, that offer information both at the application and network level of a typical browsing session. The BigFoot platform is used as the Analytics as a Service layer for both batch processing of RAW data, and for data exploration activities, using DiNoDB.
YEAR 3: H2020 IOStack Project

The IOStack project is a new project in the H2020 framework that focuses on software defined storage technologies for Big Data applications. In this context, the BigFoot project – and specifically Eurecom – has established a collaboration to enrich the Apache OpenStack modules underlying BigFoot, such that they can enjoy new data sources that were not initially considered in the BigFoot project, such as object storage and volume technology.

In addition, GirdPocket also collaborates with the IOStack project, with new use-cases that requires storage flexibility.
5 Open Source Software

In BIGFOOT, communication and dissemination of project ideas and results in terms of software deliverables to the open-source community is crucial. This Section presents the open-source projects that have been published in the first and second year of the project. Before delving into the detail of this Section, we remark that the open-source software produced by the BigFoot consortium has been studied, evaluated and validated by the OCEAN project, and appear in the list of “cloud assets” as certified for their quality.

It should be noted that not all software deliverables produced in the BigFoot project will be released as open-source software. For example, the Time Series Library (described in deliverables D.2.3.2, and discussed in this document) constitutes a new software product for GRIDPOCKET, and will not be released publicly. If not otherwise stated, the standard license used for open and free source projects is the Apache Software Foundation License, V.2.0.

We note that achieving the main milestones that were set for BIGFOOT in the DoW is the result of an incremental process: interim software deliverables – that is early releases, auxiliary software components, bug fixes – also require an appropriate dissemination strategy. In BigFoot, the publication rate of interim software deliverables, takes the following form:

- The creation of new repositories is done in a private space on BitBucket
- Early-stage software projects are released as public repositories on BitBucket
- Public software deliverables are published on GitHub

Open-source software deliverables

In the first year of the project, a number of software deliverable reached the stage of being entitled to be publicly visible on GitHub. In summary, the following projects were released:

- OSMEF [WP2, WP5]: This open-source project provides extensible tools to perform measurements (from the networking point of view, in the first release) of the performance achieved by a private cloud infrastructure, both for bare-metal and for virtual machines. This

\[^{1}\] Each project is associated to the relevant work-packages in which the implementation work has been executed.
tool is important in BIGFOOT, as it allows to understand the impact of the research work done in WP5, and as it provides an upper bound to the application-level performance available for data-intensive parallel programs running in the infrastructure. In the future, the plan is to extend the software to cover other kinds of I/O measurements (e.g. disk I/O) and many types of traffic patterns, that can be specified through a simple language.

- **datascience [WP2, WP3, WP6]**: This open-source project corresponds to the dissemination of knowledge gained and produced throughout the BIGFOOT project. It constitutes the base material that was used for the first Ph.D. school on Big Data and Applications, and touches themes ranging from the design of scalable algorithms to their execution on a parallel execution framework such as Hadoop.

- **Hadoop-Log-Tools [WP2, WP3]**: This open-source project provides a series of simple tools to examine the performance of the Hadoop batch processing engine. It has been created in the context of WP3, to analyze the behavior of job schedulers and to compare the performance of a new scheduling component produced in WP3. More generally, this tool helps in benchmarking the effects of the optimization work done in WP3, with a fine-grained view on the internals of the Hadoop parallel execution engine.

- **HFSP [WP3]**: This open-source project materializes as a “contrib” module that can be easily and seamlessly plugged in Hadoop, contributing with a new job scheduler designed and implemented in WP3. The preliminary release of HFSP is complemented by a research paper that will appear at the IEEE BigData 2013 conference. HFSP has been evaluated on the BIGFOOT platform, using Hadoop-Log-Tools, SWIM, pig-mix and other benchmarking utilities described in D.2.4.1

- **spark-kmeans [WP2]**: This open-source project constitutes a first attempt in going beyond Hadoop MapReduce. It implements a multi-dimensional version of the \(k\)-means algorithm to analyze SPAM email data (provided by SYM). The programming model is that of MapReduce, but the execution framework is SPARK, which provides several improvements over Hadoop for running iterative algorithms.

- **SWIM [WP2]**: SWIM is an open-source project originally created by the AMP lab at UC Berkeley, in cooperation with Cloudera. The
overall objective of the project is to provide a tool to generate synthetic workloads by ingesting real traces of production cluster operating Hadoop. In the context of the BIGFOOT project, SWIM has been forked and several patches have been offered to fix bugs and to extend its functionality. SWIM has been used to benchmark HFSP, and will be used for future works on scheduling performance.

In the second year of the project, the number of software deliverables dramatically increased. In the following, we present new software packages, and gloss over software packages that have been released in the first year, but updated in the second year (for example, OSMEF and HFSP). An exception is “schedsim”, which was initially released only on the Bitbucket account of BigFoot, and now is also available on the Github account of the project, after a refactoring of its source code. One additional note is in order: an important contribution from the BigFoot project to the open-source community has been that of Apache OpenStack Sahara and the Spark plug-in. This software deliverable is currently shipped with the “Ice-house” release (and subsequent releases of course) of OpenStack. As such, it is not visible on the BigFoot project Github page: instead, it can be downloaded from the legacy Apache OpenStack repositories. The following software components have been delivered in the second year as open-source projects:

- **schedsim [WP2, WP3]**: this is an open-source project that provides a simulation environment to study scheduling disciplines in general, that is, it simplifies the Hadoop execution engine to eliminate the complications of its internal resource management (the presence of slots). This software has been (and will be) heavily used during the the design of HFSP, and constitutes a theoretical benchmarking utility to quickly test scheduling disciplines.

- **TreeLib [WP2, WP3]**: this is a software library that implements Decision Trees (for both regression and classification purposes) and Random Forests, written in MapReduce, to be executed on an Apache Spark cluster. This software has been used in a collaboration between GridPocket and Eurecom (on power load forecasting) which will be presented at IEEE BigData 2014, and in a number of collaborations with SWISSCOM, AMADEUS and SeCludIT.

- **kNN graphs [WP2]**: this is an open-source project that provides an algorithm to build k nearest neighbors graphs out of textual data. It has been done as a collaboration between Symantec, Eurecom and the
Royal Military Academy in Belgium. The software has been used to understand the relationship between SPAM emails, with the ultimate goal of being able to cluster/classify emails and compare performance to MrTriage.

- **PIG ROLLUP [WP3]**: this is an open-source project that is currently under evaluation by the Apache Pig open-source community. It implements (as described in deliverables D.3.2 and D.3.3) a work-sharing primitive to optimize the execution of the ROLLUP operator, which is very common in data summarization workloads.

- **MR-TRIAGE [WP2]**: currently, the approval from Symantec to promote this software deliverable is under scrutiny, but it is highly-likely that the software will be indeed released as an open-source project. As described in D.2.3.2, MrTriage is a full-fledged analysis tool to understand and study SPAM campaigns.

In the third year of the project the following software deliverables have been produced:

- **Sahara and Sahara Image Elements**: this is the consolidation effort in Year 3, concerning the research and development activities of WP5. Such software has been accepted upstream in the current releases of Apache OpenStack.

- **DiNoDB**: this is the consolidation effort of the joint work between WP3 and WP3 on efficient and scalable queries over RAW data. Part of the project is currently awaiting for approval for its disclosure. This means that currently, the DiNoDB service is available in the list of BigFoot Analytics services as a binary code.

- **AUTO-ROLLUP**: this is the consolidation of the work-sharing optimization on large scale aggregation queries for batch engines. It has been accepted and integrated in Apache Pig.

- **MR-TRIAGE and TSL**: these are the two main software components developed in WP2, which have been consolidated in the third year of the project. MR-TRIAGE is opens source, whereas TSL is a product with proprietary code.

Note that in the third year of the project there have been minor releases of additional software components and tools that are reported in the “Software Components” tab of the BigFoot website. We do not provide additional
details here as these tools are simple utilities to interact with, for example, Apache OpenStack, and which are not deemed to be flagship contributions to the project.

**Early-stage software projects**

In Phase 1 of the project, a number of software deliverables were delivered as public repositories on BitBucket. The visibility of these projects, albeit being public, has been contained to achieve the stage of maturity required to appear on GitHub. The home page of the **BIGFOOT** project on BitBucket is: [https://bitbucket.org/bigfootproject](https://bitbucket.org/bigfootproject)

In summary, the following projects were released:

- **MumakNext** [WP2]: the original Mumak open-source project is integrated to Hadoop. It constitues an emulation environment that can be used to assess the performance impact of internal components developed for Hadoop. MumakNext is an open-source project forked from Mumak, which addresses several bugs, that extends the support for new Hadoop job schedulers, and that extends the components used to accept input configuration and data to run an emulation. Although the project is stable, it has not yet been published on GitHub because Hadoop v2 recently exposed a new emulator environment which could replace Mumak.

- **openstack docs** [WP5]: this is a public wiki that reports the experience gained in setting up an operational (and tuned) Infrastructure as a Service platform, based on the popular Apache Open Stack cloud operating system. It can be used to reproduce precisely the **BIGFOOT** platform, and has raised interest from the industrial community.

During the third year, while continuing to work on improvements in the previously released software packages, a new method of accessing **BIGFOOT** has been proposed as “Download BigFoot”. A new web sub-page has been created in order guide future **BIGFOOT** users through all major steps including download, installation, configuration and running of BigFoot Analytics as a Service (see Figure 7).

In order to enable simplified installation, the entire **BIGFOOT** system has been encapsulated within a virtual machine image. This image can be downloaded and directly executed on a OpenStack VM cluster.

The configuration and deployments process is explained on the web page (see Figure 8).
Deliverable 6.2 BIGFOOT Version 3.0

Figure 7: BigFoot Analytics download page

Figure 8: BigFoot Analytics as a Service online configuration instructions

FP7-ICT-ICT-2011.1.2 Call 8 Project No. 317858
6 Conclusion

The first twelve months of work in WP6 have been devoted to setting quality dissemination, communication and collaboration channels and building relations with external partners. Work in this WP has been organized so that all project members could take their dissemination and communication initiatives, while being supported by BigFoot partners and infrastructure.

The second year of work in WP6 has been focused on building strong collaborations with external projects and partners, as well as leveraging on internal collaboration between the BigFoot consortium partners, which results in several publication accepted by highly ranked scientific conferences.

The final year consolidates the position of BigFoot project as one of the leading initiatives in the area of Analytics as a Service. This consolidation has materialized through several high quality scientific publication, multiple demonstration and talks, and above all by release of the all-in-one, easy to use BigFoot system images.

Additionally, during the third year, an important effort has been made for large promotion of the project’s vision and results. This promotion has been achieved thanks to publication and recorded talks, and production of a series of promotional videos. The successful termination of the project and availability of deliverables is being announced by a press release.

Contributions. Below we summarize the research contributions and the value added to the BigFoot project for WP6.

- **Scientific publications**
  - We have published 39 scientific papers (14 in the first year, 13 in the second year and 12 in the last year) based on the research covering all technical work packages of BigFoot.

- **Communications**
  - We have put in place a public web site proposing various BigFoot related contents to large public (project description, deliverables, open source software, twitter exchanges and more).
  - We have presented BigFoot activities at 40 scientific and industrial event (19 in the first year and 11 in the second year and 11 in the third year) including seminars, tutorials, talks and conferences.
Deliverable 6.2

BIGFOOT
Version 3.0

– BigFoot as a whole, and selected components have been demonstrated in 15 different occasions (2 in the first and second year of the project, and 11 in the third year of the project).

• Collaboration

– We have organized a workshop on Distributed Cloud Computing in Germany, and a Ph.D. school on Big Data.
– We have contributed to 2 European concertation meetings, and to the FUI 2014 conference
– We have established 4 new collaboration initiatives with international companies and one collaboration with an academic/military institution.
– We have established concrete collaborations with two EU-funded projects, one from the FP7 framework and one from the H2020 framework.

• Open Source Software

– We have published 12 different software tools (10 software tools in the first year, 7 in the second year, and a complete ready to use virtual machine image package in the final year), projects and tutorials under the Apache Software Foundation License V2.0.