Project Acronym:	**LeanBigData**
Project Title:	**Ultra-Scalable and Ultra-Efficient Integrated and Visual Big Data Analytics**
Project Number:	**619606**
Instrument:	**STREP**
Call Identifier:	**ICT-2013-11**

D9.3 First Workshop Report

Work Package:	*WP9 – Exploitation, Industrial Awareness, Dissemination*	
Due Date:		31/01/2017
Submission Date:		31/01/2017
Start Date of Project:		01/02/2014
Duration of Project:		36 Months
Organisation Responsible for Deliverable:		ICCS/NTUA
Version:		1.0
Status:		Final
Author(s):	Vrettos Moulos, Achilleas Marinakis	ICCS/NTUA
		CA
		SyncLab
		PT
		UPM
		LeanXcale
		FORTH
		Intel
		INESC
		Atos
Reviewer(s):	Ricardo Jiménez	
Nature:	☒ R – Report ☐ P – Prototype ☐ D – Demonstrator ☐ O - Other	
Dissemination level:	☒ PU - Public ☐ CO - Confidential, only for members of the consortium (including the Commission) ☐ RE - Restricted to a group specified by the consortium (including the Commission Services)	
Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		

Revision history			
Version	**Date**	**Modified by**	**Comments**
0.1	23/12/2016	Vrettos Moulos, Achilleas Marinakis (ICCS/NTUA)	Initial Version
1.0	30/01/2017	Vrettos Moulos, Achilleas Marinakis (ICCS/NTUA)	Final Version

Executive Summary

LeanBigData dissemination plan included the organization of two public workshops to create awareness about the project results. This document reports about the first public workshop titled “Towards Convergence of Big Data, SQL, NoSQL, NewSQL, Data streaming/CEP, OLTP and OLAP - DataDiversityConvergence 2016”, which was held on 23 – 25 April 2016 in Rome, Italy.

In total 11 papers were presented, covering different aspects of LeanBigData project.

Table of Contents

Index of Figures

Index of Tables

Abbreviations and Acronyms

API	Application Programming Interface
CEBR	Centre of Economics and Business Research
CEP	Complex Event Processing
D	Deliverable
DFS	Distributed File Systems
DoW	Description of Work
EU	European Union
HDFS	Hadoop Distributed File System
ICT	Information and Communications Technology
IoT	Internet of Things
PaaS	Platform as a Service
RDMA	Remote Direct Memory Access
SDD	SEPA Direct Debit
SEPA	Single Euro Payments Area
SN	Social Networking
SOA	Service Oriented Architecture
SQL	Simple Query Language
WP	Work Package

1. Introduction

LeanBigData is an ultra-scalable and ultra-efficient big data platform integrating in one product the three main big data technologies: a novel transactional NoSQL key-value data store, a distributed complex event processing (CEP) system, and a distributed SQL database. The platform is designed to achieve scalability in a very efficient way avoiding the inefficiencies and delays introduced by current Extract-Transform-Load-based (ETL) approaches. Currently, one of the main issues in data management at enterprises and other organizations is the fact that databases are either operational (OLTP-OnLine Transactional Processing) or analytical (OLAP-OnLine Analytical Processing). This leads to a separation of the management of the operational data performed at operational databases, and the management of analytical queries performed at analytical databases or data warehouses. This separation results in having to copy the data periodically from the operational database into the data warehouse. This copy process is termed Extract-Transform-Load (ETL). ETLs are estimated to consume 75-80% of the budget for business analytics.

LeanBigData solves this issue in data management by bringing a database, LeanXcale, with the two capabilities, operational and analytical.

Another aspect in which LeanBigData innovates lies in the efficiency of the transactional processing and the storage engine. The transactional processing has been re-architected and re-implemented to be an order of magnitude more efficient than the initial version at the beginning of the project. A new storage engine, KiVi, has been architected and implemented from scratch. It is based on a new data structure to be efficient both for range queries and updates.

Another main innovation brought by LeanBigData is in the area of data streaming. Here, the goal has been to produce an efficient scalable distributed complex event processing engine

LeanBigData platform is equipped with a visualization subsystem able to report incremental visualization of results of long analytical queries and with an advanced anomaly detection and root cause analysis module. The visualization subsystem also supports efficient manipulations of visualizations and query results through hand gestures.

LeanBigData is driven by four real-world use cases with real big data sets and needs that go beyond current technology in different markets.

1. The first market is cloud data centre management and the use case focuses on cloud data centre monitoring.
2. The second one is the financial/banking market and the use case focuses on electronic alignment of direct debit transactions.
3. The third one is social networks and the use case focuses on social analytics.
4. The fourth one is telecommunications and the use case focuses on optimization of targeted advertisement distribution on web pages.

Among the 11 papers that were presented during the first LeanBigData workshop, some of them are directly related to the aforementioned use cases. Indicatively:

- Work in paper with title “Direct Debit Frauds: A Novel Detection Approach” and authors Gaetano Papale, Luigi Sgaglione, Gianfranco Cerullo, Giovanni Mazzeo, Pasquale Starace and Ferdinando Campanile starting from real SDDs data, presented an analysis of emerging attack patterns against Direct Debit transactions, it has categorized them in misuse cases and defined four Detection Criteria, in the context of the financial/banking use case
- Work in paper with title “3D Vizualization of Large Scale Data Centres” and authors Giannis Drossis, Chryssi Birliraki, Nikolaos Patsiouras, George Margetis and Constantine

Stephanidis deals with the monitoring and the intuitive display of existing data centers' information, using their actual layout, in order to inform data center experts about the servers' current state and assist navigation in actual space, in the context of the cloud data center management use case

- Work in paper with title "Big IoT and Social Networking Data for Smart Cities Algorithmic Improvements on Big Data Analysis in the Context of RADICAL City Applications" and authors Evangelos Psomakelis, Fotis Aisopos, Antonios Litke, Konstantinos Tserpes, Magdalini Kardara and Pablo Martínez Campo successfully combines citizens' posts retrieved through smartphone applications and social networks in the context of smart city applications, to produce a testbed for applying multiple analysis functionalities and techniques, in the context of the social networks use case

The project is divided into nine work packages. This deliverable belongs to work package 9 and is the result of task 9.3 "Public workshops".

2. First Public Workshop

2.1. Date and Place of the Workshop

The first public LeanBigData workshop was organized on 27th month of the project, April 2016.

The goal for choosing the place and the dates of the workshop was to find an easily reachable place in which a big event (possible conference) would be organized. The workshop should co-located with this event in order to attract more people from the academia and the industry for better dissemination outcome.

The workshop's name was "Towards Convergence of Big Data, SQL, NoSQL, NewSQL, Data streaming/CEP, OLTP and OLAP - DataDiversityConvergence 2016" and took place on 23 – 25 April 2016 in Rome, Italy [1].

It was held in conjunction with the 6th International Conference on Cloud Computing and Services Science - CLOSER 2016 [2] (Figure 1) and the International Conference of Internet of Things and Big Data - IoTBD 2016 [3] (Figure 2).

Figure 1: CLOSER 2016 poster

Figure 2: IoTBD 2016 poster

2.2. Workshop's Committees

The workshop's organizing committee, including the project's technical coordinator and the project's manager, is presented in Table 1:

Name	Affiliation
Dr. Ricardo Jimenez-Peris	LeanXcale, Spain
Prof. Marta Patiño-Martinez	UPM, Spain
Dr. Patrick Valduriez	INRIA, France
Prof. Theodora Varvarigou	NTUA, Greece

Table 1: Workshop's organizing committee

The workshop's program committee (Table 2) has been selected by the project's manager:

Name	Affiliation
Dr. Boyan Kolev	INRIA, France
Dr. José Pereira	INESC TEC & UMinho, Portugal
Dr. Valerio Vianello	UPM, Spain
Eng. Pavlos Kranas	NTUA, Greece
Eng. Sotiris Stamokostas	NTUA, Greece

Table 2: Workshop's program committee

2.3. Content of the Workshop

Papers related to the LeanBigData project and to Data Management technologies were presented to the workshop.

The workshop's deadlines for the Call for Papers were:

Deadline Description	Deadline Date
1st Call For Papers	Thu 12 Nov 2015
2nd Call For Papers	Tue 22 Dec 2015
Last Call For Papers	Mon 11 Jan 2016
Paper Submission	Fri 22 Jan 2016
Reviews Assignment	Mon 25 Jan 2016
Reviewing Process Deadline	Fri 05 Feb 2016
Author Notification	Wed 10 Feb 2016
Camera Ready and Registration	Wed 24 Feb 2016
Cancelation Policy Deadline	Mon 14 Mar 2016

Table 3: Call for papers deadlines

Since the workshop was organized in cooperation with INSTICC Events, all the papers were submitted through PRIMORIS - Event Management System [4], which supported all stages of the submission, reviewing and registration processes. The workshop chairs were in charge of the reviewing process ensuring that all papers get at least two reviews, following a double-blind process.

During the workshop 11 papers have been presented in 3 sessions of a full day. Table 4 depicts the workshop's schedule:

No.	Workshop Session	Paper Title	Authors
1	April 24 Session 1 (09:00 - 10:30)	3D Vizualization of Large Scale Data Centres	Giannis Drossis, Chryssi Birliraki, Nikolaos Patsiouras, GeorgeMargetis and Constantine Stephanidis
2		Big IoT and Social Networking Data for Smart Cities Algorithmic Improvements on Big Data Analysis in the Context of RADICAL City Applications	Evangelos Psomakelis, Fotis Aisopos, Antonios Litke, Konstantinos Tserpes, Magdalini Kardara and Pablo Martínez Campo
3		PaaS-CEP - A Query Language for Complex Event Processing and Databases	Ricardo Jiménez-Peris, Valerio Vianello and Marta Patiño-Martinez
4	April 24 Session 2 (10:45 - 12:15)	KVFS: An HDFS Library over NoSQL Databases	Emmanouil Pavlidakis, Stelios Mavridis, Giorgos Saloustros and Angelos Bilas
5		Data Collection Framework - A Flexible and Efficient Tool for Heterogeneous Data Acquisition	Luigi Sgaglione, Gaetano Papale, Giovanni Mazzeo, Gianfranco Cerullo, Pasquale Starace and Ferdinando Campanile
6		Direct Debit Frauds: A Novel Detection Approach	Gaetano Papale, Luigi Sgaglione, Gianfranco Cerullo, Giovanni Mazzeo, Pasquale Starace and Ferdinando Campanile
7	April 24 Session 3 (16:00 - 18:30)	Reducing Data Transfer in Parallel Processing of SQL Window Functions	Fábio Coelho, José Pereira, Ricardo Vilaça and Rui Oliveira

8		Design of an RDMA Communication Middleware for Asynchronous Shuffling in Analytical Processing	Rui C. Gonçalves, José Pereira and Ricardo Jimenez-Peris
9		Design and Implementation of the CloudMdsQL Multistore System	Boyan Kolev, Carlyna Bondiombouy, Oleksandra Levchenko, Patrick Valduriez, Ricardo Jimenez-Peris, Raquel Pau and José Pereira
10		Towards Quantifiable Eventual Consistency	Francisco Maia, Miguel Matos and Fábio Coelho
11		Towards Performance Prediction in Massive Scale Datastores	Francisco Cruz, Fábio Coelho and Rui Oliveira

Table 4: Workshop's schedule

2.4. Workshop's Dissemination

The dissemination of the workshop was performed through various popular means such as:

- a web site was designed and properly linked for the disseminating purpose: http://closer.scitevents.org/DataDiversityConvergence.aspx?y=2016 (Figure 3)

Figure 3: Dissemination through the workshop's website

- a relevant announcement was posted in the project's website: http://leanbigdata.eu/mcm_news/first-leanbigdata-public-workshop/

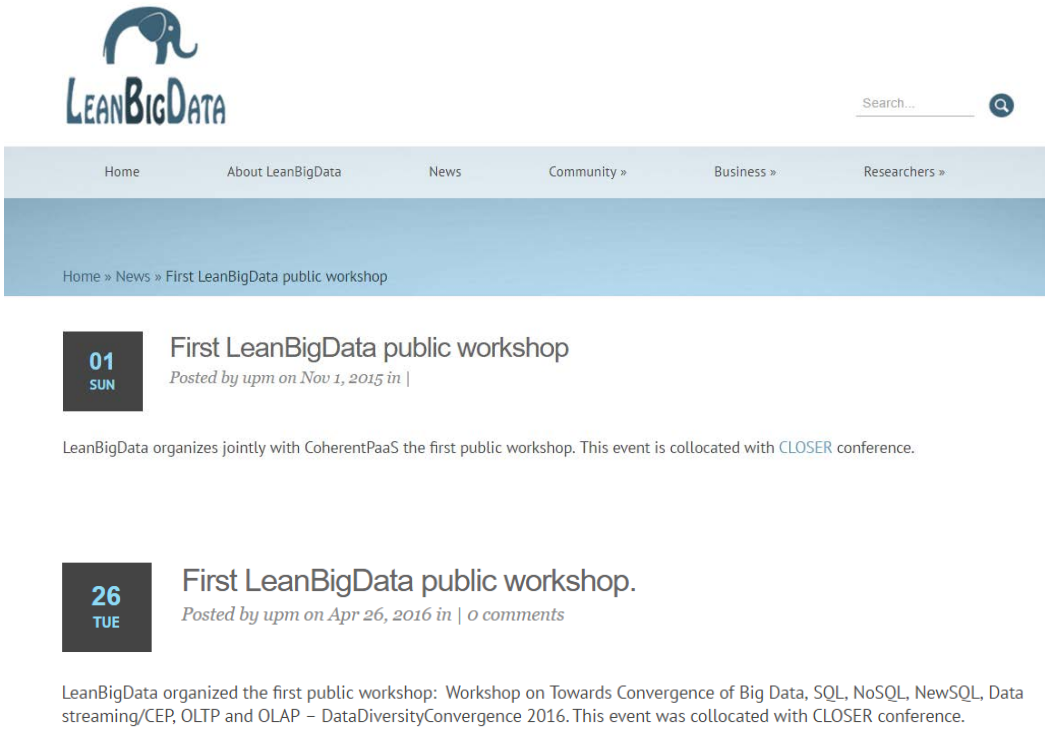

Figure 4: Dissemination through the project's website

- distributing the call for papers and advertising the workshop in the DBWorld [5] mailing list. The relevant message that was posted from the consortium can be found in the following hyperlink: https://research.cs.wisc.edu/dbworld/messages/2016-01/1452317459.html

3. References

[1]. DataDiversityConvergence 2016, http://closer.scitevents.org/DataDiversityConvergence.aspx?y=2016

[2]. Closer 2016, http://closer.scitevents.org/?y=2016

[3]. IoTBD 2016, http://www.iotbd.org/?y=2016

[4]. PRIMORIS - Event Management System, http://www.insticc.org/Primoris

[5]. DBWorld, https://research.cs.wisc.edu/dbworld/

Annex A. Paper Abstracts

List of the 11 papers abstracts which were presented at the LeanBigData First Public Workshop:

1. Title: **3D Visualization of Large Scale Data Centres**
   Abstract: This paper reports on ongoing work regarding interactive 3D visualization of large scale data centres in the context of Big Data and data centre infrastructure management. The proposed approach renders a virtual area of real data centres preserving the actual arrangement of their servers and visualizes their current state while it notifies users for potential server anomalies. The visualization includes several condition indicators, updated in real time, as well as a color-coding scheme for the current servers' condition referring to a scale from normal to critical. Furthermore, the system supports on demand exploration of an individual server providing detailed information about its condition, for a specific timespan, combining historical analysis of previous values and the prediction of potential future state. Additionally, natural interaction through hand-gestures is supported for 3D navigation and item selection, based on a computer-vision approach.
2. Title: **Big IoT and Social Networking Data for Smart Cities - Algorithmic Improvements on Big Data Analysis in the Context of RADICAL City Applications**
   Abstract: In this paper we present a SOA (Service Oriented Architecture)-based platform, enabling the retrieval and analysis of big datasets stemming from social networking (SN) sites and Internet of Things (IoT) devices, collected by smart city applications and socially-aware data aggregation services. A large set of city applications in the areas of Participating Urbanism, Augmented Reality and Sound-Mapping throughout participating cities is being applied, resulting into produced sets of millions of user-generated events and online SN reports fed into the RADICAL platform. Moreover, we study the application of data analytics such as sentiment analysis to the combined IoT and SN data saved into an SQL database, further investigating algorithmic and configurations to minimize delays in dataset processing and results retrieval.
3. Title: **PaaS-CEP - A Query Language for Complex Event Processing and Databases**
   Abstract: Nowadays many applications must process events at a very high rate. These events are processed on the fly, without being stored. Complex Event Processing technology (CEP) is used to implement such applications. Some of the CEP systems, like Apache Storm the most popular CEPs, lack a query language and operators to program queries as done in traditional relational databases. This paper presents PaaS-CEP, a CEP language that provides a SQL-like language to program queries for CEP and its integration with data stores (database or key-value store). Our current implementation is done on top of Apache Storm however, the CEP language can be used with any CEP. The paper describes the architecture of the PaaS-CEP, its query language and the algebraic operators. The paper also details the integration of the CEP with traditional data-stores that allows the correlation of live streaming data with the stored data.
4. Title: **KVFS: An HDFS Library over NoSQL Databases**
   Abstract: Recently, NoSQL stores, such as HBase, have gained acceptance and popularity due to their ability to scale-out and perform queries over large amounts of data. NoSQL stores typically arrange data in tables of (key, value) pairs and support few simple operations: get, insert, delete, and scan. Despite its simplicity, this API has proven to be extremely powerful. Nowadays most data analytics frameworks utilize distributed file systems (DFS) for storing and accessing data. HDFS has emerged as the most popular choice due to its scalability. In this paper we explore how popular NoSQL stores, such as HBase, can provide an HDFS scale-out file system abstraction. We show how we can design an HDFS compliant filesystem on top a key-value store. We implement our design as a user-space library (KVFS) providing an HDFS filesystem over an HBase key-value store. KVFS is designed to run Hadoop style analytics such as MapReduce, Hive, Pig and Mahout over NoSQL stores without the use of HDFS. We perform a preliminary evaluation of KVFS against a native HDFS setup using DFSIO with varying number of threads. Our

results show that the approach of providing a filesystem API over a key-value store is a promising direction: Read and write throughput of KVFS and HDFS, for big and small datasets, is identical. Both HDFS and KVFS throughput is limited by the network for small datasets and from the device I/O for bigger datasets.

5. Title: **Data Collection Framework - A Flexible and Efficient Tool for Heterogeneous Data Acquisition**
   Abstract: The data collection for eventual analysis is an old concept that today receives a revisited interest due to the emerging of new research trend such Big Data. Furthermore, considering that a current market trend is to provide integrated solution to achieve multiple purposes (such as ISOC, SIEM, CEP, etc.), the data became very heterogeneous. In this paper a flexible and efficient solution about the data collection of heterogeneous data is presented, describing the approach used to collect heterogeneous data and the additional features (pre-processing) provided with it.
6. Title: **Direct Debit Frauds: A Novel Detection Approach**
   Abstract: Single Euro Payments Area (SEPA) is an initiative of the European banking industry aiming at making all electronic payments across the Euro area as easy as domestic payments currently are. One of the payment schemes defined by the SEPA mandate is the SEPA Direct Debit (SDD) that allows a creditor (biller) to collect directly funds from a debtor's (payer's) account. It is apparent that the use of this standard scheme facilitates the access to new markets by enterprises and public administrations and allows for a substantial cost reduction. However, the other side of the coin is represented by the security issues concerning this type of electronic payments. A study conducted by Center of Economics and Business Research (CEBR) of Britain showed that from 2006 to 2010 the Direct Debit frauds have increased of 288%. In this paper a comprehensive analysis of real SDD data provided by the EU FP7 LeanBigData project is performed. The results of this data analysis will conduct to define emerging attack patterns that can be execute against SDD and the related effective detection criteria. All the work aims at inspire the design of a security system supporting analysts to detect Direct Debit frauds.
7. Title: **Reducing Data Transfer in Parallel Processing of SQL Window Functions**
   Abstract: Window functions are a sub-class of analytical operators that allow data to be handled in a derived view of a given relation, while taking into account their neighbouring tuples. We propose a technique that can be used in the parallel execution of this operator when data is naturally partitioned. The proposed method benefits the cases where the required partitioning is not the natural partitioning employed. Preliminary evaluation shows that we are able to limit data transfer among parallel workers to 14% of the registered transfer when using a naive approach.
8. Title: **Design of an RDMA Communication Middleware for Asynchronous Shuffling in Analytical**
   Abstract: A key component in a distributed parallel analytical processing engine is shuffling, the distribution of data to multiple nodes such that the computation can be done in parallel. In this paper we describe the initial design of a communication middleware to support asynchronous shuffling of data among multiple processes on a distributed memory environment. The proposed middleware relies on RDMA (Remote Direct Memory Access) operations to transfer data, and provides basic operations to send and queue data on remote machines, and to retrieve this queued data. Preliminary results show that the RDMA-based middleware can provide a 75% reduction on communication costs, when compared with a traditional sockets implementation.
9. Title: **Design and Implementation of the CloudMdsQL Multistore System**
   Abstract: The blooming of different cloud data management infrastructures has turned multistore systems to a major topic in the nowadays cloud landscape. In this paper, we give an overview of the design of a Cloud Multidatastore Query Language (CloudMdsQL), and the implementation of its query engine. CloudMdsQL is a functional SQL-like language, capable of querying multiple heterogeneous data stores (relational, NoSQL, HDFS) within a single query that can contain embedded invocations to each data store's

native query interface. The major innovation is that a CloudMdsQL query can exploit the full power of local data stores, by simply allowing some local data store native queries (e.g. a breadth-first search query against a graph database) to be called as functions, and at the same time be optimized.

10. Title: **Towards Quantifiable Eventual Consistency**
    Abstract: In the pursuit of highly available systems, storage systems began offering eventually consistent data models. These models are suitable for a number of applications but not applicable for all. In this paper we discuss a system that can offer a eventually consistent data model but can also, when needed, offer a strong consistent one.
11. Title: **Towards Performance Prediction in Massive Scale Datastores**
    Abstract: Buffer caching mechanisms are paramount to improve the performance of today's massive scale NoSQL databases. In this work, we show that in fact there is a direct and univocal relationship between the resource usage and the cache hit ratio in NoSQL databases. In addition, this relationship can be leveraged to build a mechanism that is able to estimate resource usage of the nodes composing the NoSQL cluster.