HiPerDNO

High Performance Computing Technologies for Smart Distribution Network Operation

FP7 - 248135

Project coordinator: Dr Gareth Taylor, BU

Consortium Members: BU, EF, IBM ISRAEL, University of Oxford, EENL, UNION FENOSA, INDRA, GTD, KORONA, EG, Fraunhofer IWES

Document Title	Report of First Performance Test of the Data Mining Platform
Document Identifier	HiPerDNO/2011/D1.3.2
Version	0.1
Work package number	WP1
Sub-Work package Number	WP1.3
Distribution	Public
Reporting consortium member	IBM
Internal reviewer & review date	Yehuda Naveh 01/08/2011

First Performance Test of the Data Mining Platform

Summary

Future electricity distribution network operators (DNO) with mass deployment of network equipment sensors will generate vast amounts of data, which requires scalable data mining techniques in order to turn the data into actionable information. To meet these challenges DNOs can benefit from the use of techniques recently developed to cost-effectively solve large scale data mining problems using high performance computing platforms. This in turn, will bring operations and maintenance more online moving the industry from *reactive* to *proactive* operations, which can lead to actions that will improve electrical grid reliability.

A key ingredient in the development process of large scale data mining applications is a scalable data mining platform. A *scalable distributed data mining platform* is a software library that includes a collection of fundamentals algorithms in machine learning (e.g. clustering, classification, dimension reduction, regression analysis and pattern mining), which are implemented in a parallel programming paradigm in order to run on top of commodity computing clusters. These scalable data mining platforms are indispensable to the data miner as they aim to assist building large-scale intelligent systems easier and faster. To this end, this report surveys two leading public open source mining platforms coming from two different approaches using well established categories for evaluating data mining software within a smart gird environment.

In essence, our findings are summarized as follows. On one hand, most matured statistical software packages, including GNU R, are originally geared towards deep analytics with vast variety of functionality, but do not scale easily to large data. Furthermore, both leading parallel packages for parallel computing with GNU R, and deliver good performance support a spectrum of functionality, but in terms of usability the development process requires significant expertise in parallel programming. On the other hand, scalable data mining platforms, which are built on top of scalable database management system, including Apache Mahout, scale to large datasets and provide hooks for user-defined functions and procedures, but they do not deliver the rich analytic functionality found in statistical packages. Nevertheless, due to the massive shift in the data mining community towards Apache Hadoop in choosing to implement scalable data mining platforms, we expect to significant improvements in Apache Mahout's functionality and performance in the near future as both Apache Hadoop and Apache Mahout mature.

Document Information

Project Number	FP7 - 248135		Acronym	HiPerDNO
Full Title	First Performance Test of the Data Mining Platform			
Project URL	http://www.hiperdno.eu			
Document URL	N/A			
Deliverable Number	D1.3.2	**Title**	Report of First Performance Test of the Data Mining Platform	
Work Package Number	WP1	**Title**	Research and Development of HPC and Communications for Large Scale Data Processing in Distribution Networks	

Delivery	**Work Plan Date**	1st August 2011	**Actual Date**	1st August 2011
Status	Version 1.0			
Nature	Prototype ☐ Internal Report ☒ External Report ☐ Dissemination ☐			
Dissemination Level	Public ☒ Consortium ☐			
Author(s) (Partners)	IBM			
Lead Author	**Name**	Yaacov Fernandess	**E-mail**	yaacov@il.ibm.com
	Partner	IBM	**Phone**	97248296324
Abstract	A key ingredient in the development process of large scale data mining applications for power grids is a scalable data mining platform. A *scalable distributed data mining platform* is a software library that includes a collection of fundamentals algorithms in machine learning, which are implemented in a parallel programming paradigm in order to run on top of commodity computing clusters. These scalable data mining platforms are indispensable to the data miner as they aim to assist building large-scale intelligent systems easier and faster. To this end, this report reviews two leading public and open source scalable data mining platforms coming from two different approaches using well established categories for evaluating scalable data mining software within a smart gird environment.			
Keywords	high performance computing, machine learning, data mining, parallel computing, computer cluster, benchmarks			

Table of Contents

1 Introduction

Future electricity distribution network operators (DNO) with mass deployment of network equipment sensors will generate vast amounts of data, which requires analysis in order to turn the data into actionable information. To meet these challenges DNOs can benefit from the use of techniques recently developed to cost-effectively solve large scale computational problems in areas such as Biology, Finance and Web Services. In such systems, increased access to ubiquitous sensing and the web has resulted in an explosion in the size of data mining and machine learning tasks, which in turn, driven the growing demand for *scalable* implementations of machine learning algorithms on very large datasets (ranging from 100s of GBs to TBs of data). In the meantime, physical and economic limitations have forced computer architecture towards parallelism and away from exponential frequency scaling. In general, parallel computing - often called distributed computers - deals with hardware and software for computation in which many calculations are carried out simultaneously. There are different types of existing architectures and technologies for parallel computing but in this report we focus on *commodity computing cluster*. A computer cluster is a group of shared individual computers, linked by high-speed communications in a local area network, and incorporating system software which provides an integrated parallel processing environment for applications with the capability to divide processing among the nodes in the cluster.

In order to benefit from current and future trends in parallel computing there are several attempts at building scalable distributed data mining platforms on top of commodity computing clusters. A *scalable distributed data mining platform* is a parallel computing application that includes a collection of fundamentals algorithms in machine learning. That is the algorithm's computation is distributed on large set of cluster's nodes rather than processed on a single core machine. In general, deploying parallel computing application on commodity computing cluster becomes increasingly challenging as the number of jobs, users or compute nodes increases. A resource management tool can alleviate such issues. Resource manager systems are software applications to submit, control and monitor jobs in cluster computing environments. These applications work on the operating system level and mostly directly monitor the communication protocols and provide an execution environment for the scalable data mining platform. Therefore, they are independent from the data mining platform. Nonetheless, we address compatibility of candidates platforms with the specific HPC platform, which is suggested in HiPerDNO Deliverable 1.2.1 and 1.2.2, in deliverable HiPerDNO Deliverable 3.1.1.

Scalable data mining platforms are indispensable to the data analyst as they aim to assist building large-scale intelligent systems easier and faster. The advantage is clear, building on open source developed and public software projects will allow DNOs to focus their development effort on the core functionality rather than on complementary component in nature, such as, machine learning library. To this end, this report reviews two leading open source and free scalable data mining platforms coming from two different approaches using well established categories for evaluating scalable data mining software within a smart gird environment. The reminder of this report is structured as follows. In Section 2, we review machine learning applications for power grids. Section 3 introduces our selection criteria for scalable data mining platform and in Section 4 we demonstrate it on Apache Mahout and Parallel GNU R. Last, in Section 5 we give our conclusions.

2 Data Mining for Power Grids

According to [R9] surveys conducted in Europe and North America reliability will be a key issue as electrical grids transform to smart grids throughout the next several decades, and grid maintenance will become even more critical than it is currently. As grid parts are replaced gradually and as smart components are added, the old components, including cables, switches, sensors, etc., will still need to be maintained. Maintaining a large grid that is a mix of new and old components is more difficult than managing a new grid. Thus, the key to making smart grid components effective is to use data mining and statistical machine learning for preventive maintenance, and in turn preventing cascade failures. More concretely, the electrical grid data, which contains monitoring information of the distribution network, can be transformed into machine learning and data mining models that aim to predict grid reliability and assisting with maintenance actions. To this end, data mining application, which runs on top of the high performance computing platform, will incorporate machine learning algorithms on very large datasets that are implemented by scalable data mining library. Consequently, electrical smart grid will bring operations and maintenance more online moving the industry from *reactive* to *proactive* operations, which in turn, can lead to actions that will improve electrical grid reliability.

Data mining applications for power grids include various applications such as the prediction of power security breaches, forecasting, power system operation, control and maintenance, and classification of power system disturbances. The goal these applications is to demonstrate that data collected by electrical utilities can be used to create statistical models for proactive maintenance, to exemplify how this can be accomplished through state-of-the–art data mining techniques, and show how DNOs can be most effective in building predictions and decision support application. To name but a few examples: Rudin et al. [R9] present a seminal work on the New York City power grid. The work introduces new methodologies for maintaining and ranking assets in the smart grid, in the form of a general process for failure prediction that can be specialized for individual applications. In [R10] the authors applied multiple regression and neural network to recognize partial discharge in electrical transformers. Last, [R2] introduces a model for forecasting long-term electricity load for the Egyptian power grid and [R5] introduces an intelligent system to monitor the power voltage and to detect power quality disturbances.

3 Categories for Evaluating Scalable Data Mining Platform for Power Grids

Currently, scalable data mining platforms are expensive and selection of the wrong platform can be costly in many ways. The cost of selecting an improper scalable data mining platform for a particular application is even more costly in terms of personnel resources, development time, and the potential for acting on spurious results. Moreover, evaluating scalable data mining platforms is not simply a matter of selecting the best tool for all purposes. Instead a DNO must consider the platforms with respect to their particular environment, and analysis needs.

Das et al. [R6] distinguish between two main approaches adopted in the implementation of scalable data mining platform. The first approach includes statistical platforms such as GNU R, IBM SPSS, SAS or Matlab. Each of these platforms provides a comprehensive environment for statistical computation, including a concise statistical language, well tested libraries of statistical algorithms for data exploration and modelling, and visualization facilities. Most of these statistical platforms were originally designed to target the moderately-sized datasets commonly found in other areas of statistical practice. For large scale datasets these platforms usually include specialized software packages for parallelizing. In parallel to the development of statistical platforms, the database community has developed a variety of large-scale data management systems that can handle huge amount of data. Examples include traditional enterprise data warehouses and newer systems based on Apache Hadoop [R12]. In terms of analytics, however, such systems have been limited primarily to simple computation, they do not deliver rich analytics functionality found in statistical platforms.

To better understand and to evaluate the different scalable data mining platforms that are available, it is important to have an overview of existing technologies for parallel computing. For future developments and wide adoption, it is important for all packages to use standardized programming paradigm, and to establish platform-independent solutions.

In this report we review two leading open source scalable data mining platforms coming from the two different approaches given above. The first approach is represented by The Parallel GNU R [R18], which is leading statistical platform, which includes parallelism packages based on Open MPI [R14]. The second is represented by Apache Mahout [R11], which is scalable data mining platform implemented using framework called Apache Hadoop [R12]. Experience and research [R1] suggests three major categories of criteria for evaluating scalable data mining platforms: functionality, usability, scalability and performance, which are addressed next.

3.1 Functionality

This category focuses in the inclusion of a variety of capabilities, techniques, and methodologies for data mining. Software functionality helps assess how well the tool will adapt to different data mining problem domains, and in turn, allows organisations to focus their development effort on the core functionality rather than on complementary component in nature.

3.2 Usability

This category concerns with the accommodation of different levels and types of users without loss of functionality or usefulness. In this category we are concern with following questions: Is the tool easy to learn and use? How well suited is the tool for its target user type? How easy is the tool for analysts to use? How well does it focus on a variety of domains? In addition, a good tool will provide meaningful diagnostics to help debug problems and improve the output.

3.3 Scalability and Performance

In scalability we mean the effect that an increase in the size of the training set has on the computational performance of a data mining algorithm. That is, we measure how well the algorithm scales to large data sets. However, scalability and performance of the different parallel computing platforms can differ due to a number of reasons. Among these reasons, design and efficiency of the implementation as well the technology and hardware are likely the dominant factors. Nonetheless, in this category we review qualitative aspects of the scalable data mining platform's ability to easily handle well known benchmarks in data mining, which are commonly used in data mining and are likely to be used in the future, thereby achieving a realistic representation of the existing applications exist in HiPerDNO project.

4 Evaluation of Existing Scalable Data Mining Platforms

4.1 Parallel GNU R

GNU R is an open-source programming language and software environment for statistical computing and graphics. The GNU R language has become a de facto standard among statisticians for developing statistical software, and is widely used for statistical software development and data analysis. GNU R was originally created by R. Ihaka and R. Gentleman in 1993 and is now being developed by the R Development Core Team [R18]. The core GNU R installation provides the language interpreter and many statistical and modelling functions. In addition, GNU R is highly extensible through the use of packages. Packages are libraries for specific functions or specific areas of study, frequently created by GNU R users and distributed under suitable licenses. A large number of packages are available at the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/.

Providing software for parallel or high performance computing (HPC) with GNU R was not a development goal. GNU R is designed to target the moderate-size dataset commonly found in areas of statistical practice, and to operate on a single server. Nonetheless, nine different GNU R packages for cluster-based parallel computing are available at CRAN. An excellent survey of existing parallel computing packages for GNU R is given in [R4].

The existing approaches to parallelizing GNU R can be classified by their degree of abstraction. These range from low-level message passing to task and data-parallel processing to high-level automatic parallelization. According to [R4], two packages stand out as particularly suited to general use on computer clusters, namely Rmpi [R19] and SNOW [R22], which are both based on MPI [R14] among other parallelism techniques. MPI is a standardized and portable message-passing system designed to function on a wide variety of parallel computers. The MPI interface is meant to provide essential virtual topology, synchronization, and communication functionality between a set of processes in a language-independent way. OpenMPI [R14] is a project combining technologies and resources from several other MPI projects with the stated aim of building the best MPI library available.

The package Rmpi is a wrapper to MPI, providing a GNU R interface to low-level MPI functions. In this way, the GNU R user does not need to know details of the MPI implementations. It requires that MPI is installed, and runs under popular MPI implementations. The Rmpi package includes scripts to launch GNU R instances at the slaves from the master, and provides several R-specific functions for error-handling and to report errors form the workers to the manager. The package SNOW (Simple Network of Workstations) supports simple parallel computing in GNU R. SNOW is a higher-level task and data-parallel computing systems for parallelizing GNU R that built on top of a low level

message-passing package (e.g. via Rmpi), and therefore, easier to use. SNOW provides functionality to spawn a cluster, to distribute values across the cluster, and to apply in parallel a given function to a large set of alternative arguments. SNOW can be used to implement task parallelism (arguments are tasks) or data parallelism (arguments are data).

4.1.1 Functionality

GNU R provides a wide variety of statistical and graphical techniques. Currently, there are more than 3000 packages available in the CRAN package repository. As such, GNU R is ideally suited to the many challenging tasks associated with data mining. In fact, according to Rexer's Annual Data Miner Survey in 2010, GNU R has become the data mining tool used by more data miners (43%) than any other. For an example, RWeka[R21] a GNU R package that provides interface to Weka [R17]. Weka is a collection of machine learning algorithms for data mining tasks written in Java, containing tools for data pre-processing, classification, regression, clustering, association rules, and visualization. However, these packages were designed to target the moderate-size dataset commonly found in areas of statistical practice, and to operate on a single server.

Nonetheless, the GNU R community is very active in improving the scalability of GNU R and there are dozens of approaches that aim at parallelizing GNU R across computational cluster. As before, the main drivers for this work are increased dataset sizes and the increasing demands of scientific research and high-performance computing. A complete and comprehensive list of software resources that are useful for high-performance computing (HPC) with R can be found at the high-performance and parallel computing CRAN task [R23].

4.1.2 Usability

Usability describes how easily software can be deployed to achieve a particular goal. For computer clusters, the Rmpi and SNOW packages cover the full functionality of MPI communication standard and provide an essentially complete API to this standard. The existence of online tutorials for the packages Rmpi and SNOW improves their learning curve and they are well suited for experienced target user. However, direct use of message-passing systems requires significant expertise in parallel programming. In addition, in order to correct and fix errors, it is important that functions provided by a package help the user to identify sources of error, which are included in both packages.

To summarize, the flexibility of the R package system allows integration of many different technologies to address various practical data mining questions. According to [R4], the authors stress that whichever technologies emerge, GNU R should be well-positioned to take advantage of them. However, the complexity working with the communication API at a

relatively low level makes Rmpi and SNOW package somewhat more difficult for researchers with primarily statistical interests to become familiar with.

4.1.3 Scalability and Performance

Performance of the different parallel computing packages can differ due to a number of reasons. Among these reasons, design and efficiency of the implementation as well the technology and hardware are likely the dominant factors. Nonetheless, several works attempt to evaluate the scalability and performance of parallel GNU R for computer cluster, which we review next.

In [R4], Schmidberger et al. design a benchmark to evaluate the performance of the cluster packages in GNU R. The benchmark comprises three different components: Sending data from the manager to all workers, distributing a list of data from the manager to the workers and a classic parallel computation example from the numerical analysis literature, the integration of a three dimensional function. The authors conclude that MPI appears to be emerging as a de-facto standard for parallel computing in GNU R and is the best supported communication protocol. Moreover, the package SNOW is the fastest solutions for parallel computing with GNU R. But if the ratio of communication and calculation time is good all packages have similar performance.

Another interesting example is Ricardo [R6]. Ricardo is part of the eXtreme Analytics Platform (XAP) project at the IBM Almaden Research Centre, and rests on a decomposition of data-analysis algorithms into parts executed by GNU R statistical platform and parts handled by the Apache Hadoop data management system. This decomposition attempts to minimize the transfer of data across system boundaries. According to test performance conducted SNOW appeared to be too low-level for conveniently implementing scalable deep analytics. This assertion was based on experience in using the SNOW package to implement the computation of the latent-factor model for recommendation system. Although the experiments indicated that the performance of Apache Hadoop is suboptimal compared to SNOW, the SNOW implementation required 50% more code than the implementation in Ricardo, while at the same time being much more complicated and error prone. Additional features such as scalability to a large number of nodes, fault tolerance and elasticity were infeasible to implement using GNU R and SNOW that they were dropped from the implementation. As both packages require a “tightly” connected cluster, where fault tolerance, redundancy and elasticity are not provided. This restriction limits scalability to relatively small clusters (i.e. less than a hundred).

4.2 Apache Mahout

Apache Mahout is a new open source library by the Apache Software Foundation (ASF) with the primary goal of developing scalable machine-learning algorithms that are free to use under the Apache license. Apache Mahout began life in 2008 as a subproject of Apache's Lucene project [R13], which provides the well-known open-source search engine of the same name. As of April 2010, Apache Mahout has become a top-level Apache project in its own right. While Mahout is, in theory, a project open to implementations of all kinds of machine learning techniques, it is in practice a project that focuses on several key areas of machine learning, such as, clustering, classification, dimension reduction, regression analysis and pattern mining.

As before, Apache Mahout aims to be the machine learning library of choice when the data to be processed is very large, perhaps far too large for a single machine. To this end, the core machine learning algorithms are implemented on top of Apache Hadoop framework. Apache Hadoop is a framework for running applications on computing cluster built of commodity hardware. The Apache Hadoop framework transparently provides applications both reliability and data motion. Apache Hadoop implements a computational paradigm named Map/Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system (HDFS) that stores data on the computing nodes, providing very high aggregate bandwidth across the cluster. Both Map/Reduce and the distributed file system are designed so that node failures are automatically handled by the framework.

In the last years we have witnessed a massive shift in the data mining community towards Apache Hadoop in choosing a parallel programming paradigm in order to implement distributed data mining platforms, both in the academia and industry. In fact, most large-scale data mining libraries, including Apache Mahout, adopted Apache Hadoop as a generic parallel programming paradigm for large clusters of machines to name a few examples, [R3, R6, R8, and R11]. Consider, as an example, IBM Parallel Machine Learning Toolbox [R15], which is based on MPI parallel programming. At the time of writing the project proposal, IBM Parallel Machine Learning Toolbox was part of IBM Alpha Works and was included as possible candidate platform for HiPerDNO. However, since then IBM Parallel Machine Learning Toolbox was not adopted as leading technology. On the contrary, currently, IBM stopped supporting the MPI approach and started developing large-scale data mining platform based on Apache Hadoop, which is similar to Apache Mahout, called SystemML [R8].

4.2.1 Functionality

Several approaches to machine learning are used to solve problems. This section focuses on the two most commonly used ones in real applications, supervised and unsupervised learning. Supervised learning is the machine learning task of inferring a function from supervised training data. Many algorithms are used to create supervised learners, and the most common are integrated in Apache Mahout, such as, naive Bayes classifiers, random forest, support vector machine and logistic regression. Unsupervised learning is tasked with making sense of data without any labeled examples of what is correct or incorrect. It is most commonly used for clustering similar input into logical groups. It also can be used to reduce the number of dimensions in a data set in order to focus on only the most useful attributes, or to detect trends. Common approaches to unsupervised learning, which are integrated into Apache Mahout, include k-Means, hierarchical clustering, and spectral clustering. In addition, Apache Mahout includes algorithms for pattern mining, evolutionary and dimension reduction algorithms. In essence, Apache Mahout incubates a number of techniques and algorithms, many still in development or in an experimental phase. Nonetheless, at this early stage in the project's life, three core themes are evident: classification, clustering, and collaborative filtering from which one can construct a customized large-scale intelligent system.

4.2.2 Usability

The existence of online tutorials and documentation for the Apache Mahout improves its learning curve and it's well suited for experienced target user. Furthermore, the goal of Apache Mahout is to build a vibrant, responsive, diverse community to facilitate discussions not only on the project itself but also on potential use cases. This in turn, has turned Apache Mahout to a leading open source scalable data mining platform and it's have been adopted to assist building large-scale intelligent systems in various organizations, which run on top commodity computing clusters with thousands of nodes. For example, Yahoo! Mail uses Apache Mahout's Frequent Pattern Set Mining. NewsCred uses Mahout to generate clusters of news articles and to surface the important stories of the day. Foursquare uses Apache Mahout to help develop predictive analytics, 365Media uses Apache Mahout's Classification algorithms in its real-time system named UPTIME and 365Media/Social. Last, Dicode [R16] project, which is also FP7 EU project, which aims to facilitate and augment collaboration and decision making in data-intensive and cognitively-complex settings. To do so, it exploits and builds on the most prominent high-performance computing paradigms and large data processing technologies - such as cluster and cloud computing, Apache Hadoop and Mahout – to meaningfully search, analyze and aggregate data existing in diverse, extremely large, and rapidly evolving sources.

4.2.3 Scalability and Performance

Apache Mahout aims to be the machine learning library of choice when the data to be processed is very large, perhaps far too large for a moderate cluster size. To this end, the core machine learning algorithms are implemented on top of Apache Hadoop framework. That is, the implementation of these algorithms use the map/reduce paradigm as generic parallel programming paradigm along with HDFS, which is a distributed, scalable, and portable file system for the Apache Hadoop framework. Apache Mahout is designed to be enterprise-ready; it's designed for performance, scalability and flexibility. More concretely, Apache Mahout inherits its performance, scalability, reliability and security features from the underlying infrastructure of Apache Hadoop framework.

Apache Mahout is early in the software development cycle. Hence, there are some functions with significant test performance and some that are brand new. However, every release makes existing library functions mature and with improved performance while other functions will be just integrated. For an example, for the clustering algorithms integrated in Apache Mahout there several benchmarks, which are tested on Amazon EC2 [R24]. These benchmarks are publicly reusable dataset, which are being used by the Apache Mahout community to monitor the performance improvements in each release of cluster algorithms. In addition, for the mature recommendation system incorporated in Apache Mahout there are extensive benchmarks which are reported in the project wiki [R25]. To summarize, although experiments indicated that the performance of Mahout is still suboptimal to exiting solution, we expect significant performance improvements in the future as this technology matures.

5 Conclusions and Recommendations

To better understand and to evaluate the different scalable data mining platforms that are available, it is important to have an overview of existing approaches and open standards for parallel computing. To this end, this report reviewed two leading scalable data mining platforms, Apache Mahout and Parallel GNU R, which are based on different approaches and are implemented using a generic open source parallel programming paradigm. In order to evaluate these platforms we outline three major categories for evaluating scalable data mining software within a smart gird environment.

To conclude, at this point in time, on one hand, most matured statistical software packages, including GNU R, are originally geared towards deep analytics with vast variety of functionality, but do not scale easily to large cluster size. Furthermore, both Rmpi and SNOW support a spectrum of functionality for parallel computing with GNU R, and deliver good performance, but in terms of usability the development process requires significant expertise in parallel programming. On the other hand, scalable data mining platforms, which are built on top of scalable database management system, including Apache Mahout, scale to huge datasets and provide hooks for user-defined functions and procedures, but they do not deliver the rich analytic functionality found in statistical packages. Nevertheless, due to the massive shift in the data mining community towards Apache Hadoop in choosing to implement scalable data mining platforms, we expect to significant improvements in Apache Mahout's functionality and performance in the near future as both Apache Hadoop and Apache Mahout mature.

6 References

[R1] B. Carey and C. Marjaniemi. "A Methodology for Evaluating and Selecting Data Mining Software." In HICSS, 1999.

[R2] H.K. Mohamed, S.M. El-Debeiky, H.M. Mahmoud and K.M El Destawy. "Data Mining for Electrical Load Forecasting In Egyptian Electrical Network". In ICCES, 2006.

[R3] C.-T. Chu, S. K. Kim, Y.-A. Lin, Y. Yu, G. R. Bradski, A. Y. Ng and K. Olukotun. "Map-reduce for machine learning on multicore." In NIPS, 2006.

[R4] M. Schmidberger, M. Morgan, D. Eddelbuettel, H. Yu, L. Tierney and U. Mansmann. "State of the art in parallel computing with R". In Journal of Statistical Software, 2009.

[R5] W. Huaying, L. Jingbo and S. Xiufa. "A novel intelligent system for analysis and recognition of power quality disturbance signal." In CDC, 2009.

[R6] S. Das, Y. Sismanis, K. Beyer, R. Gemulla, P. Haas, and J. McPherson. "Ricardo: Integrating r and hadoop." In SIGMOD, 2010.

[R7] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. "GraphLab: A New Parallel Framework for Machine Learning." In UAI, 2010.

[R8] A. Ghoting, R. Krishnamurthy, E. Pednault, B. Reinwald, V. Sindhwani, S. Tatikonda, Y. Tian, and S. Vaithyanathan, "SystemML: Declarative Machine Learning on MapReduce." In ICDE, 2011.

[R9] C. Rudin, D. Waltz, R. N. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. Gross, B. Huang, S. Ierome, D. Isaac, A. Kressner, R. J. Passonneau, A. Radeva, L. Wu. "Machine Learning for the New York City Power Grid." In TPAMI, 2011.

[R10] W. Yu. "A New Method for Partical Discharge Pattern Recognition of Electrical Transformers." In ICMTMA, 2011.

[R11] Apache Mahout, "Apache mahout", http://mahout.apache.org/

[R12] Apache Hadoop, "Apache hadoop", http://hadoop.apache.org/

[R13] Apache Lucene, "Apache Lucene", http://lucene.apache.org/

[R14] Open MPI, http://www.open-mpi.org/

[R15] IBM Parallel Machine Learning Toolbox, http://www.alphaworks.ibm.com/tech/pml

[R16] Dicode, http://dicode-project.eu/

[R17] Weka, http://www.cs.waikato.ac.nz/ml/weka/

[R18] GNU R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, URL http://www.R-project.org/.

[R19] Rmpi: http://www.stats.uwo.ca/faculty/yu/Rmpi/

[R20] R and Data Mining: http://www.rdatamining.com/

[R21] RWeka http://cran.r-project.org/web/packages/RWeka/index.html

[R22] SNOW http://cran.r-project.org/web/packages/snow/index.html

[R23] HPC with R http://cran.r-project.org/web/views/HighPerformanceComputing.html

[R24] Amazon EC2 http://aws.amazon.com/ec2/

[R25] Mahout Benchmarks https://cwiki.apache.org/MAHOUT/mahout-benchmarks.html