Developing Scalable Smart Grid Infrastructure to Enable Secure Transmission System Control

EP/K006487/1

UK PI: Prof Gareth Taylor (BU)

China PI: Prof Yong-Hua Song (THU)

Consortium UK Members: Brunel University (BU), Alstom Grid Ltd (AGL), Intel Corporation (IC), National Grid (NG)

Consortium China Members: Tsinghua University (THU), Sichuan University (SU), CEPRI

Document Title	Detailed Report on suitability of trusted cloud computing platforms at enterprise level and dedicated cluster computing platforms at an application level for scalable data processing and mining
Document Identifier	
Version	Version 1.0
Work package number	WP2
Task Number(s)	T2.1 & T2.2
Distribution	Public
Reporting consortium member	BU, THU, SU and CEPRI
Internal reviewer & review date	Prof G A Taylor 28/2/2014

Document Information

EPSRC Project Number	EP/K006487/1	**China Funding Council**	NSFC
Full Title	Detailed Report on suitability of trusted cloud computing platforms at enterprise level and dedicated cluster computing platforms at an application level for scalable data processing and mining		
Project URL	http://www.brunel.ac.uk/sed/ece/research/bips/epsrc-smart-grids		
Document URL	N/A		
Report Number	T2.1&2.2	**Title**	Experimental feedback on proposed algorithms and sensors positioning (delivery: M35) N/A
Work Package Number	WP2	**Title**	Development and deployment of scalable secure high computational tools and infrastructure

Date of Delivery	**Work Plan**	M18	**Actual**	M18
Status	Version 1.0			
Nature	Internal Report			
Dissemination Level	Consortium			
Author(s) (Partners)	G A Taylor (BU), M Li (BU), M Khan (BU), M Jane (BU) Z Huang (BU)			
Lead Author	**Name**	Prof G A Taylor	**E-mail**	Gareth.Taylor@brunel.ac.uk
	Partner	BU	**Phone**	+44 (0) 1895 266610
Abstract	As the next generation of power system, smart grid brings more and more newly developed devices and techniques into the power system field. Big data analytics now becomes one of the important challenges in smart grid. New solutions are needed with regard to data mining and processing that are suitable for the scalable data of smart grid. In this report, a review on trusted cloud computing platforms and dedicated cluster computing platforms is given. The application of the trusted cloud computing platform and the dedicated cluster computing platform, Hadoop MapReduce, for the big data analytics in power system is also investigated to evaluate their suitability in smart grid data processing and mining. Some classical data mining technique are reviewed for dedicated cluster computing platform.			
Keywords	Cloud computing, Hadoop MapReduce, Data mining			

Contents

Abbreviations

EA	Evolutionary Algorithm
GA	Genetic Algorithm
GEP	Gene Expression Programming
GP	Genetic Programming
HDFS	Hadoop Distributed File System
IaaS	Infrastructure as a Service
LAN	Local Area Network
PaaS	Platform as a Service
SaaS	Software as a Service

1. Introduction

With the development of smart grid, the complexity level of data generated from power systems is increased significantly. More and more newly developed devices and techniques are brought into the current power system. Big data analytics is now one of the important challenges in the power system. The data mining and processing works in next generation power system need new solutions which are suitable for the scalable data of smart grid.

This report provides a review on trusted cloud computing platforms and dedicated cluster computing platforms. The application of the trusted cloud computing platform and the dedicated cluster computing platform, Hadoop MapReduce, for the big data analytics in power systems is investigated to evaluate their suitability in smart grid data processing and mining.

The rest of this report is organized as follows. Section 2 describes the big data analytics in the power system. Section 3 represents the trusted cloud computing platform for smart grid. Section 4 introduces the Hadoop MapReduce platform and data mining algorithm for the data processing and mining in smart grid.

2. Big Data Analytics

The next generation power system smart grid is designed to achieve the following targets-- steady availability of power, energy sustainability, environmental protection, prevention of large-scale failures, Optimized Operational Expenditure (OPEX) of power production and distribution, and reduced future Capital Expenditure (CAPEX) for thermal generators and transmission networks [1]. The analysis of the extreme larger data set generated from the smart grid has become a very popular research field. Study [2] outlines several research topics related to asset management, operation planning, real-time monitoring and fault detection/protection that present new opportunities. The analytic of dynamic energy management and power distribution systems are also involved [3][4].

Generally speaking, compared with a conventional data set the data generated from smart gird has three new characteristics: high velocity, high volume, and high variety. Smart grid needs an efficient data processing and mining platform to meet the new challenge of these three characteristics. Cloud computing technology and Hadoop MapReduce are investigated in this project in order to provide a suitable data processing and mining platform.

3. Trusted Cloud Computing

3.1 Trusted Cloud Computing

Cloud computing is a technology designed to provide scalable, virtual, shared, metered and on-demand computing resources across the Internet or Local Area Network (LAN) to small, medium and large enterprises. These resources are delivered through the following services: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [9][10]. Although cloud computing is an attractive technology, it continues to be a subject of uncertainty among end users as there is no way to evaluate the trustworthiness of its services which are most times managed by “faceless” providers across a wider, sometimes “unknown” or unclearly marked, geographical area. Trust has always been an important aspect of any computer systems including cloud computing. No standard definition exists, but trust in the cloud implies that the origin and identity of a resource or storage system can be verified and its integrity assured.

3.2 Trusted Cloud Computing and Power System

In order to address the management problem of different front intelligence devices and to process the huge amount of data, cloud computing is applied in many areas of the power system. It can be applied for time-varying computational purposes including forecasting and planning, locational marginal pricing, topology/parameter error detection, and contingency analysis. It also can be applied to supply support for reliable data storage, ease of access and the open market of utility-oriented computing services [5]. The security problem of cloud computing is also considered for power systems. Study [7] proposes a secure cloud computing-based framework, a smart frame for big data information management in smart grids. Within such a framework a hierarchical structure of cloud computing centres is built to provide different types of computing services for information management and big data analysis. In order to address the critical security problem, a solution based on identity-based encryption, signature and proxy re-encryption has been developed. Study [8] presents a use case illustrating how the integration of trusted computing technologies into an available cloud infrastructure, Eucalyptus. It allows the security-critical energy industry to exploit the flexibility and potential economical benefits of the cloud computing paradigm for their business-critical applications.

In this project, our group (M-J. Sule, etc.) also attempted to enhance the security of cloud computing [13]-[15]. The IDS engine runs within a vm and provides a measure of the one-to-many (tree) direct trust of other applications within the vm [13]. At the application level, the data-colouring is introduced to maintain integrity and availability of data with minimal processing overhead [14]. Trust modelling in cloud computing can provide some level of control, transparency and attestation for the cloud end user [15].

4. Dedicated Cluster Computing Platform & Scalable Data Processing and Mining

In order to handle the scalable data processing and mining job in the power system, a dedicated computing platform and some data mining algorithms are investigated in this project. The detail of Hadoop MapReduce platform and some popular data mining algorithms are presented below.

4.1 Hadoop MapReduce

4.1.1 MapReduce

MapReduce is a programming model designed for processing and generating large data sets. With its associated implementation, users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key [16]. A MapReduce job usually splits the input data-set into independent chunks. These chunks are then processed by the map tasks with a completely parallel mechanism [17]. This type of platform provides a highly applicable computational intensive solution for the big data of smart grid by using the parallelized computational mechanism to provide a near real-time process for the scalable dataset.

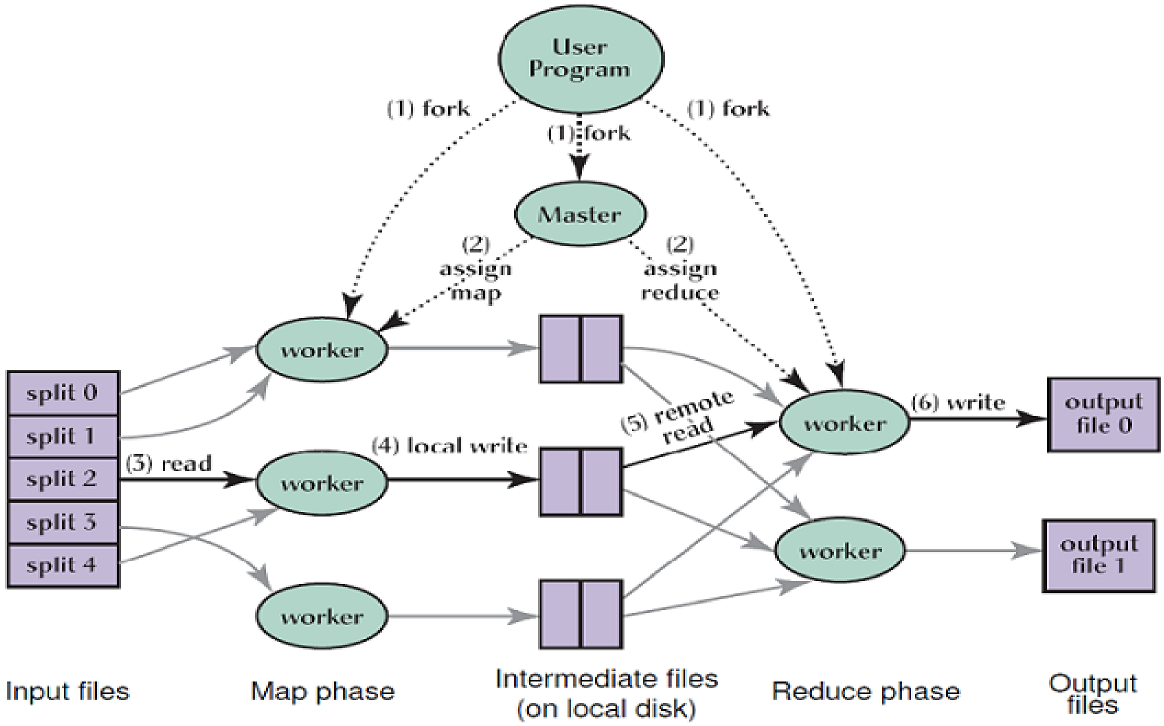

Fig. 1. MapReduce

4.1.2 HDFS

HDFS is a distributed file system that provides high-throughput access to application data. The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably and to stream those data sets at high bandwidth to user applications. By distributing storage and computation across many nodes in the network, the resources needed by the job can grow with demand. Because the very large data sets are segmented and stored separately in many nodes, the reliably of the data is also guaranteed [18][19]. This reliability and high bandwidth are two key advantages of HDFS for the data processing and mining jobs in smart grid.

4.1.3 ResourceManager

ResourceManager is an ultimate authority which controls the resources among all the applications in the system. The ResourceManager has two main components: a) Scheduler and b) ApplicationsManager. The Scheduler is designed to allocate resources to the various running applications subject to familiar constraints of capacities, queues etc. The ApplicationsManager is designed for accepting job-submissions, negotiating the first container for executing the application specific ApplicationMaster and providing the service for restarting the ApplicationMaster container on failure [20]. The resource manager provides a very flexible and stable mechanism for complex data processing jobs.

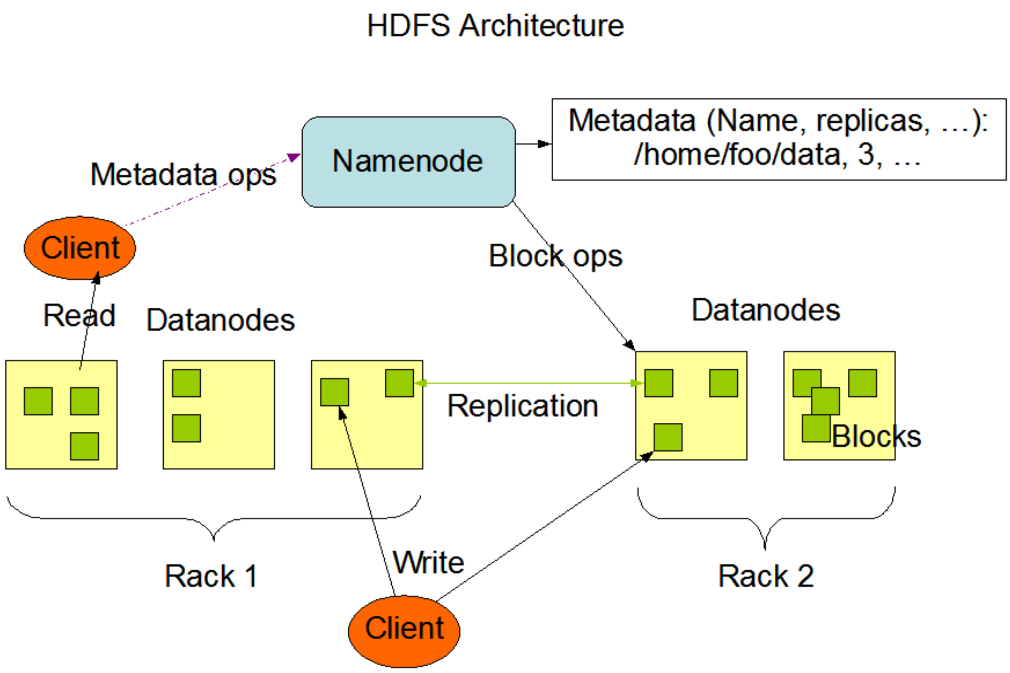

Fig. 2. HDFS Architecture

4.2 Hadoop Core Projects

In this section some Hadoop core projects are introduced and their suitability to be applied for the data mining and processing tasks is also discussed.

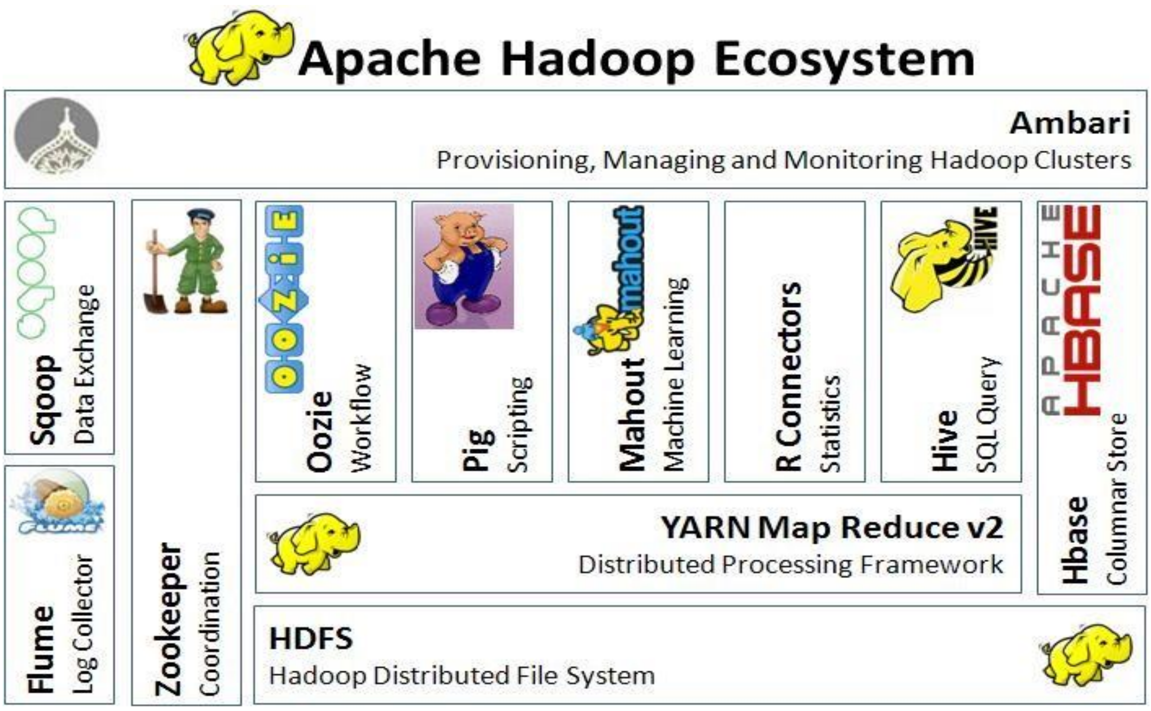

Fig. 3. Hadoop Projects

4.2.1 Apache Pig

Apache Pig is developed for analysing large data sets that consists of a high-level language for expressing data analysis programs. It also provides an infrastructure for evaluating data analysis programs. The structure of Pig programs is amenable to substantial parallelization. Very large data sets can be handled efficiently by pig programs [21]. Pig consists of two layers. A) The infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs. It handles the large-scale parallel implementations that already exist. B) The language layer which is supported by a textual language called Pig Latin. It has three key properties, which are eas of programming, optimization opportunities and extensibility.

4.2.2 Apache Hive

The Apache Hive is a data warehouse software which is designed to facilitate querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data. It uses a SQL-like language called HiveQL to query the data. HiveQL allows traditional map/reduce programmers to plug in their custom mappers and reducers. HiveQL can also take some user-defined scalar functions (UDF's), aggregations (UDAF's), and table functions (UDTF's). Although Hive does not support real-time queries, it provides scalability (more nodes can be added dynamically to the Hadoop cluster), extensibility (supports MapReduce framework and UDF/UDAF/UDTF), fault-tolerance, and loose-coupling with its input formats [22].

4.2.3 Apache Hbase

Apache Hbase is a distributed, scalable, versioned, non-relational database modeled big data store. It is designed to host very large tables (billons of rows and millions of columns) on the top of Hadoop and HDFS and achieves random, real-time access (read/write) to big data sets. Hbase provides linear and modular scalability, strictly consistent reads and writes, and automatic and configurable sharing of tables [23].

4.2.4 Apache ZooKeeper

Apache ZooKeeper is designed to develop and maintain an open-source server which provides a highly reliable distributed coordination. It is a centralized complex service which includes maintaining configuration information, naming, providing distributed synchronization, and providing group services [24].

4.2.5 Apache Sqoop

Apache Sqoop is designed to provide an efficient data transfer between the Apache Hadoop and some structured data stores [25]. With Sqoop the bulk data can be efficiently transferred between Hadoop platform and data stores such as relational databases.

4.3 Data Processing and Mining

In order to address the data processing and mining problem in smart grid, some classical data mining technologies and newly developed algorithms are introduced and discussed in this section. The suitability of parallelization of those data processing and mining algorithms are also investigated. The problems in developing algorithms for a distributed environment are discussed. The algorithms for parallelized and distributed data mining are compared and discussed.

4.3.1 Background

Data processing is a collection and manipulation of items of data to produce meaningful information. Data mining is a computational process of discovering patterns in big data [26]. Due to the increases in data sizes and algorithmic complexities, the computational power of a single core process is not sufficient to satisfy the requirements of data processing and mining. Parallelized and distributed computational models are developed to provide HPC support. Study [27] provides a performance evaluation and characterization of some parallelized scalable data mining algorithms. The algorithms of parallelized and distributed data mining include:

- For Classification problem:
  - ScalParC: an efficient and scalable variation of decision tree classification
  - Naive Bayesian classifier: comparable in performance to decision trees and exhibits high accuracy and speed when applied to large databases.
- For Clustering problem:
  - K-means: a partition-based method and is arguably the most commonly used clustering technique
  - BIRCH: a hierarchical clustering method that employs a hierarchical tree to represent the closeness of data objects
  - HOP: a typical density-based clustering method
- For Association problem:
  - Apriori: arguably the most influential ARM algorithm
  - Eclat uses a vertical database format to break the search space into small, independent, and manageable chunks

Their parallel efficiency is shown in Table 1 [27]. In the table “-” indicates “not implemented”.

4.3.2 Application in Power System

With the growth of data in the power system, data mining techniques were developed dramatically in the past years. Study [28] provides a review on 18 data mining applications in power systems. These successful power system applications include statistical analyses solutions and artificial intelligence solutions. Similar to these classical data mining algorithms, the computational intensive problem can be solved with parallelization and distribution techniques. Can [29] provides a benchmark of data mining solution for the application in power system. In this study the author indicated that the input data size is a very important agent in achieving reasonable speedup when parallelizing using Hadoop cluster. The bigger data size needs a larger Hadoop cluster with more nodes.

In order to deal with the big data in the power system, the parallelization with Hadoop cluster is the most scalable and the most efficient way to implement data mining processes [30]. The increasing data set size can be handled by growing the number of computing nodes.

Algorithm	Execution time (s) one core	Execution time (s) four core	Execution time (s) eight core
HOP	52.7	1.92	6.06
K-means	12.9	3.9	4.96
Fuzzy K- means	146.8	3.44	5.42
BIRCH	31.7	-	-
ScalParC	110.6	3.88	5.12
Bayesian	25.1	-	-
Apriori	102.7	2.66	3.36
Eclat	81.5	-	-

Table 1. Parallel efficiency of data mining algorithm [27]

4.3.3 Statistical Analyses Solution

Statistical analyses are one of the popular data processing and mining solutions in power system. Study [31] provides a general overview of statistical learning theory. In this study some theoretical support were provided for the suitability of applying parallel computation for the statistical solution. Zhao et al. [32] also provide a parallelized K-means clustering solution with Hadoop MapReduce. The performance of parallel implementation was evaluated with speedup, scaleup and sizeup. The result demonstrates that large scalable datasets can be processed effectively with Hadoop MapReduce.

4.3.4 Artificial Intelligence Solution

Powered by the evolutionary theory, artificial intelligence algorithms also provide some good solutions for data processing and mining problems in power systems. They use an organized and structured searching mechanism to explore the solution space. This kind of organized and structured searching mechanism could benefit more from Hadoop MapReduce. The parallelized and distributed computation model can also be implemented more efficiently.

4.3.4.1 Gene Expression Programming

Gene Expression Programming (GEP) is a member of Evolutionary Algorithms (EAs). It is developed based on a similar idea to Genetic Algorithms (GA) and Genetic Programming (GP) [4]. In order to maintain and accumulate the genetic information, GEP operates a separated genotype and phenotype system to handle the representation of the candidate solution. The algorithm inherits the advantages of the linear structure of GA and of the tree structure of GP. This linear and tree combination provides a structured and flexible mechanism by searching through the space of a big data problem. Because of such structured and flexible mechanisms, GEP is a very suitable candidate to be accelerated with parallelized or distributed mechanisms in order to handle a big scalable data set.

In this project, GEP was used to optimize the performance of Hadoop MapReduce platform to process PMU data. A parallelized version of GEP was also developed to benefit from the multi-core processor.

4.3.4.2 Neural Network

Neural network is another popular artificial intelligence solution for data mining and processing problems. It uses a system of interconnected "neurons" which exchange messages between each other to represent the knowledge generated from the data set. This kind of interconnected network also can be accelerated with parallelized or distributed mechanisms. However, the correlation of each factor in the data set is not well considered. It is better to be applied on data set which contains relative loosen factor correlation. Study [33] provides an example of application on parallelizing neural network for training data on a cluster. The sub sets of data are processed on different nodes of the cluster.

4.3.4.3 Genetic Algorithm

The Genetic Algorithms (GA) are used to solve search, optimization and machine learning problems. The original GA use a fixed-length string to represent the solution. The simplest representation is a fixed string of zeros and ones (0 and 1). This kind of fixed-length structure provides a very good platform for the parallelization and distribution computation. Compared with GEP and neural network, GA can achieve the highest parallelization efficiency. However, GA does not consider the correlation of each involved factor in a system at all. Therefore, GA is more suitable for processing and mining a data set which contains independent factors.

5. Conclusion

With the development of modem information technology, the data that need to be processed in current power system become more and more large and complex. The number of data sources, the number of data formats and the speed of data gathering are increasing dramatically. In order to investigate the suitability of applying high performance computing techniques in smart gird, this report reviewed the trusted cloud computing, the dedicated cluster computing platform- Hadoop MapReduce and several data mining and processing solutions. The result demonstrates that the trusted cloud computing and Hadoop MapReduce are very good platforms for the data processing and mining tasks. The trusted cloud computing provides a secured data processing system which can handle the multiple sources (formats) of data generated from the smart grid. The statistical and artificial intelligence methods equipped with Hadoop MapReduce, such as cluster computing platforms, also provide many very efficient data mining solutions.

References

[1] S. F. Bush, S. Goel, and G. Simard, *IEEE Vision for Smart Grid Communications: 2030 and Beyond Roadmap*. 2013.

[2] M. Kezunovic, L. Xie, and S. Grijalva, “The role of big data in improving power system operation and protection,” in *Proceedings of IREP Symposium: Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid, IREP 2013*, 2013.

[3] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, “Big Data Analytics for Dynamic Energy Management in Smart Grids,” *Big Data Res.*, vol. 2, no. 3, pp. 94–101, 2015.

[4] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo, “Big data analytics in power distribution systems,” in *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2015, pp. 1–5.

[5] Dán, György, Rakesh B. Bobba, George Gross, and Roy H. Campbell. "Cloud computing for the power grid: from service composition to assured clouds." In *Proceedings of the 5th USENIX Workshop on Hot Topics in Cloud Computing. USENIX Association*. 2013.

[6] J. Baek, Q. H. Vu, J. K. Liu, X. Huang, and Y. Xiang, “A secure cloud computing based framework for big data information management of smart grid,” *IEEE Trans. Cloud Comput.*, vol. 3, no. 2, pp. 233–244, 2015.

[7] F. Ma, X. Luo, Q. Zhang, and E. Litvinov, “Cloud computing: an innovative IT paradigm to facilitate power system operations,” in *Power & Energy Society General Meeting, 2015 IEEE*, 2015, pp. 1–5.

[8] D. Wallom, M. Turilli, G. Taylor, N. Hargreaves, A. Martin, A. Raun, and A. McMoran, “myTrustedCloud: Trusted cloud infrastructure for security-critical computation and data managment,” in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, 2011, pp. 247–254.

[9] P. Mell and T. Grance, “The NIST Definition of Cloud Computing (Draft) Recommendations of the National Institute of Standards and Technology,” *Nist Spec. Publ.*, vol. 145, no. 6, p. 7, 2011.

[10] D. Sun, G. Chang, L. Sun, and X. Wang, “Surveying and analyzing security, privacy and trust issues in cloud computing environments,” *Procedia Eng.*, vol. 15, pp. 2852–2856, 2011.

[11] R. K. L. Ko, P. Jagadpramana, M. Mowbray, S. Pearson, M. Kirchberg, Q. Liang, and B. S. Lee, “TrustCloud: A framework for accountability and trust in cloud computing,” in *Services (SERVICES), 2011 IEEE World Congress on*, 2011, pp. 584–588.

[12] T. Beth, M. Borcherding, and B. Klein, “Valuation of Trust in Open Networks,” *Proc. 3rd Eur. Symp. Res. Comput. Secur.*, vol. 875, pp. 3–18, 1994.

[13] M.-J. Sule, M. Li, G. A. Taylor, and S. Furber, “Deploying trusted cloud computing for data intensive power system applications,” in *Power Engineering Conference (UPEC), 2015 50th International Universities*, 2015, pp. 1–5.

[14] M-J. Sule, M. Li and G. A. Taylor, “Securing Data Resources in Trusted Clouds” (Submitted to Future Generation Computer Systems (Elsevier) – September 2015)

[15] M-J. Sule, M. Li and G. A. Taylor, " Trust Modeling in Cloud Computing" (Submitted to IEEE SOSE Conference 29th March – 2nd, April, 2016)

[16] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[17] “MapReduce” [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html. [Accessed:03-Dec-2015]

[18] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The hadoop distributed file system,” in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, 2010, pp. 1–10.

[19] (2014, Mar.). An Introduction to HDFS High Availability [Online]. Available: http://www.cloudera.com/content/cloudera-content/clouderadocs/CDH4/latest/CDH4-High-Availability-Guide/cdh4hag_topic_2_1.html

[20] “Resource Manager” [Online]. Available: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html. [Accessed:03-Dec-2015]

[21] “Apache pig” [Online]. Available: https://pig.apache.org/ [Accessed:03-Dec-2015]

[22] “Apache hive” [Online]. Available: https://hive.apache.org/ [Accessed:03-Dec-2015]

[23] “Apache Hbase” [Online]. Available: https://hbase.apache.org/. [Accessed:03-Dec-2015]

[24] “Apache ZooKeeper” [Online]. Available: https://zookeeper.apache.org/. [Accessed:03-Dec-2015]

[25] “Apache sqoop” [Online]. Available: http://sqoop.apache.org/. [Accessed:03-Dec-2015]

[26] K. Funatsu, *New Fundamental Technologies in Data Mining*. InTech, 2011.

[27] Y. Liu, J. Pisharath, W. Liao, G. Memik, A. Choudhary, and P. Dubey, “Performance evaluation and characterization of scalable data mining algorithms,” in *16th IASTED international conference on parallel and distributed computing and systems (PDCS). MIT, Cambridge*, 2004, pp. 620–625.

[28] J. Morais, A. Klautau, C. Cardoso, and Y. Pires, *An overview of data mining techniques applied to power systems*. INTECH Open Access Publisher, 2009.

[29] C. Anil, “Benchmarking of Data Mining Techniques as Applied to Power System Analysis,” 2013.

[30] Z. A. Vale, C. Ramos, S. Ramos, and T. Pinto, “Data mining applications in power systems --Case-studies and future trends,” in *Transmission & Distribution Conference & Exposition: Asia and Pacific, 2009*, 2009, pp. 1–4.

[31] V. N. Vapnik, “An overview of statistical learning theory.,” *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[32] W. Zhao, H. Ma, and Q. He, "Parallel k-means clustering based on mapreduce," in *Cloud computing*, Springer, 2009, pp. 674–679.

[33] I. Monedero, C. Leon, J. Ropero, A. Garcia, J. M. Elena, and J. C. Montano, "Classification of Electrical Disturbances in Real Time Using Neural Networks," *IEEE Trans. Power Deliv.*, vol. 22, no. 3, pp. 1288–1296, 2007.