6th Iranian Conference on Electrical and Electronics Engineering
(ICEEE2014)
Islamic Azad University Gonabad Branch
August 19,20,21 - 2014

Application of Hadoop for Offline Analysis of Power Quality Disturbances

Nader Mollayi
Department of Computer and Information Technology
Birjand University of Technology
Birjand, Iran
mollayi@birjandut.ac.ir

Seyyed Hadi Mousavi
Department of Computer and Information Technology
Birjand University of Technology
Birjand, Iran
mousavi@birjandut.ac.ir

Farhad Nazarzadeh Dabbagh
Department of Computer and Information Technology
Birjand University of Technology
Birjand, Iran
f.nazarzadeh.dabbagh@gmail.com

***Abstract*—Identification of voltage and current disturbances is an important task in power system monitoring and protection. Due to the wide frequency range of the disturbances, a high sampling rate is necessary for offline processing of the disturbances, which leads to a large volume of data and such a processing is impractical, therefore. In this paper application of Hadoop distributed computing software, for offline processing of power quality disturbances is proposed and it is shown that this application makes such a processing possible and leads to a very cheaper system with widespread usage, compared to the power quality analyzers.**

***Keywords**- power quality, power qulaity disturbances, offline processing of power qulaity disturbances, distributed processing, big data, Hadoop*

I. INTRODUCTION

Electrical energy is the most widespread type of energy used in world today, due to its simpler transmission, distribution, conversion and usage. AC electrical power system is in use for over a century, worldwide. In this system, the best electrical supply would be a constant magnitude and frequency sinusoidal voltage waveform. However, because of the non-zero impedance of the supply system, large variety of loads and of other phenomena such as transients and outages, the reality is often different. The Power Quality of a system expresses to which degree a practical supply system resembles the ideal supply system.

In recent years, there has been an increased concern for the quality of power due to the rapid developments of power electronic devices and their widespread use in industry. These devices are major sources of power quality problems, and from the other hand, they are much more sensitive to voltage disturbances than their counterparts in the past. Utility switching and fault clearing also affects the quality of delivered power. Most of the equipment in use today is susceptible to damage or service interruption during poor power-quality events [1]. In addressing this problem, the Institute of Electrical and Electronics Engineers (IEEE) has done significant work on the definition, detection, and mitigation of power quality events, and the events are classified into 7 categories of power quality disturbances based on IEEE 1159 standard [2].

To improve electric power quality, sources and causes of disturbances must be specified before taking any mitigation action. In order to achieve this purpose, events must be detected and classified [3]. Power quality analyzer is a measurement instrument used to measure and monitor the power quality disturbances online. Due to the large number of points which must be analyzed and high price of this device, monitoring of power quality parameters is performed regularly in special points and very often by electric distribution companies with poor economic status, such as 3rd world countries. Offline processing of power quality disturbances is also very hard to perform and beyond capabilities of a personal computer due to the large volume of data which must be processed.

In this paper, offline processing of power quality disturbances based on Hadoop distributed computing software is proposed, and it is shown that this application makes such a processing possible and results in a much cheaper system with widespread usage.

II. POWER QUALITY DISTURBANCES

Mono frequency sinusoidal waveform with constant magnitude and an equal phase difference in three phases is the ideal voltage waveform in ac power systems, and deviations from these conditions are regarded as power quality disturbances. Power quality disturbances are divided into seven categories based on IEEE 1159 standard. Categories and typical duration of the disturbances based on this standard are listed in Table 1 [2]. Three types of the disturbances are depicted in Figs. 1-4 [4, 5].

It could be obviously seen that the disturbances contain a wide frequency range and various characteristics. Therefore, a high sampling frequency is required for digital processing of the disturbances, and several complex algorithms are developed for their detection and analysis.

TABLE I. CATEGORIES AND TYPICAL DURATION OF POWER QUALITY DISTURBANCES AS DEFINED BY IEEE 1159 STANDARD

Category		Subcategory	Typical duration
Transients	Impulsive	nanosecond	<50 ^{ns}
		microsecond	50 ^{ns} – 1 ^{ms}
		milisecond	>1 ^{ms}
	Oscillatory	Low frequency	0.3-50 ms
		Medium frequency	20 $^{\mu s}$
		High frequency	5 $^{\mu s}$
Short duration Variations (Sag, Swell, Interruption)		Instantaneous	0.5-30 cycles
		Momentary	30 cycles- 3 ^{s}
		Temporary	3 ^{s}-1 ^{min}
Long duration Variations			> 1 min
Voltage Imbalance			Steady State
Waveform Distortion	DC Offset		Steady State
	Harmonics (2-100th)		Steady State
	Interharmonics (0-6^{khz})		Steady State
	Notching		Steady State
	Noise (Typical Magnitude:0-1%) BroadBand Frequency Range		Steady State
Voltage Fluctuations (Frequency Range < 25 ^{Hz})			Intermittent
Power Frequency Variations			<10 ^{s}

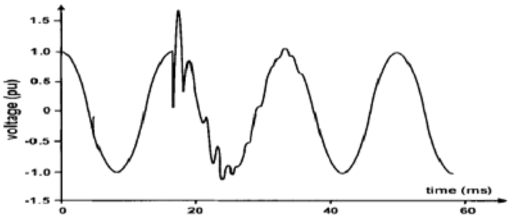

Figure 1. An Oscillatory Transient Caused by Capacitor bank switching

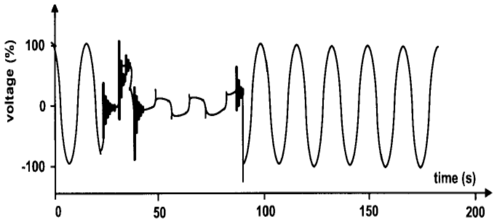

Figure 2. Voltage sag caused by single line to ground fault

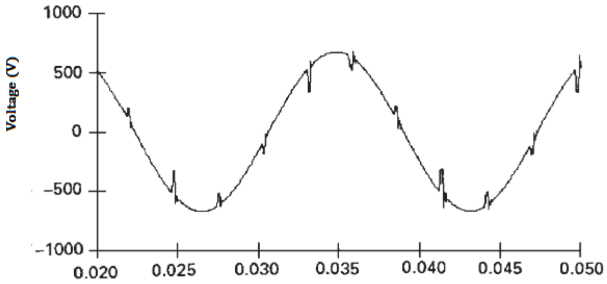

Figure 3. Voltage notching caused by a three phase converter.

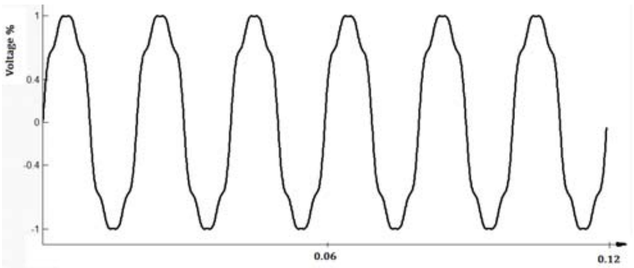

Figure 4. A harmonic polluted waveform

III. DIGITAL PROCESSING OF POWER QUALITY DISTURBANCES

In order to process an analog signal digitally, a sampling rate twice the highest frequency component of the signal to be processed, is necessary based on Nyquist-Shannon sampling theorem [6]. If we neglect high frequency voltage transients with duration of less than 20 $^{\mu s}$, or frequency contents of higher than 50 ^{kHz}, a minimum sampling frequency of 100 ^{kHz} is necessary for this purpose. However transients with higher frequency contents could also be detected from their effects in current, due to the inductive property of the power system. A higher sampling frequency may not result in better results due to the low-pass frequency response of CT and PT, also [7].

Based on this fact, a sampling rate of 2048 samples per cycle in a 50 ^{Hz} system is the minimum sampling rate necessary for digital processing of the disturbances, properly. Twelve bits of data is also necessary for an acceptable sampling precision, resulting in quantization error of less than 200^{mv} .

This sampling rate results in data volume listed in second row of table 2 row for each signal to be recorded. The total recording volume must be six times this amount since the voltage and current in three phases must be processed in a three phase system [8].

TABLE II. DATA VOLUME REQUIRED FOR RECORDING POWER QUALITY DISTURBANCES WITH 2048 SAMPLES/CYCLE IN A 50 HZ SYSTEM

Duration	1 Second	1 Minute	1 Hour	1 Day	1 Week
Data Volume for Each Signal	150 ^{Kb}	8.79 ^{Mb}	527.34 ^{Mb}	12.36 ^{Gb}	86.52 ^{Gb}
Total Data Volume	900 ^{Kb}	52.74 ^{Mb}	3.09 ^{Gb}	74.16 ^{Gb}	519.12 ^{Gb}

In order to process the data by a digital computer, any sample must be converted to floating point with 16 bits at least. Therefore the volume of data to be processed is the volume of recorded data multiplied by 4/3, resulting to the total data to be processed as listed in table 3.

TABLE III. DATA VOLUME REQUIRED FOR PROCESSING POWER QUALITY DISTURBANCES WITH 2048 SAMPLES/CYCLE IN A 50 HZ SYSTEM

Duration	1 Second	1 Minute	1 Hour	1 Day	1 Week
Total Data Volume	1200 ^{Kb}	70.32 ^{Mb}	4.12 ^{Gb}	98.88 ^{Gb}	692.16 ^{Gb}

IV. DETECTION AND CLASSIFICATION OF POWER QUALITY DISTURBANCES

Due to the wide frequency range and various characteristics of power quality disturbances, detection and classification of these disturbances is a complex task and several algorithms are developed for this purpose. Discrete Fourier Transformation (DFT) and Discrete Wavelet Transform (DWT) are traditionally used for detection of steady state disturbances and transients, respectively [9].

Discrete Fourier Transformation decomposes a signal to its frequency components and is implemented by Fast Fourier Transformation (FFT) algorithm. Discrete wavelet transformation simply decomposes a signal to approximation and details which contain the low and high frequency components of the signal, respectively. Steady state disturbances are identified via the frequency spectrum calculated by FFT and the details resulting from discrete wavelet transformation are suitable for characterization of transients, therefore [10-12]. Decomposition of A pure sinusoidal waveform, a harmonic polluted waveform and their decomposition into frequency components based on FFT are shown in Figs.5, 6. Fig.7 depicts decomposition of the sinusoidal waveform in Fig.6.A into approximation and details by DWT in one level based on db6 motherwavelet. The waveform in Fig.8.A contains an oscillatory transient. Wavelet based decomposition of this signal is shown in Figs.8.b, c. The transient forms the main pattern of the detail and its energy is 18.77 times the energy of the detail for pure sinusoidal waveform.

Therefore, disturbances are detected based on the following parameters in this paper:

1- Frequency components which are not negligible in comparison with main frequency component
2- The total energy of details.

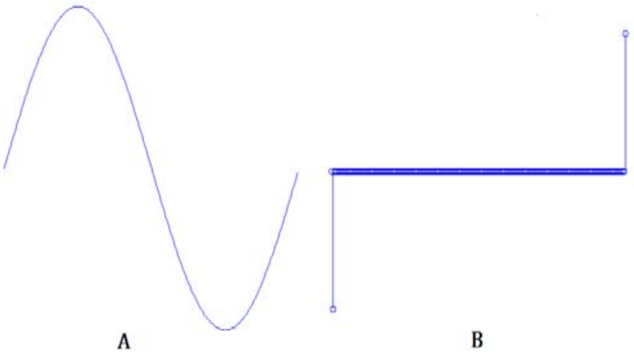

Figure 5. A) A pure sine wave B) Its frequency decomposition based on FFT

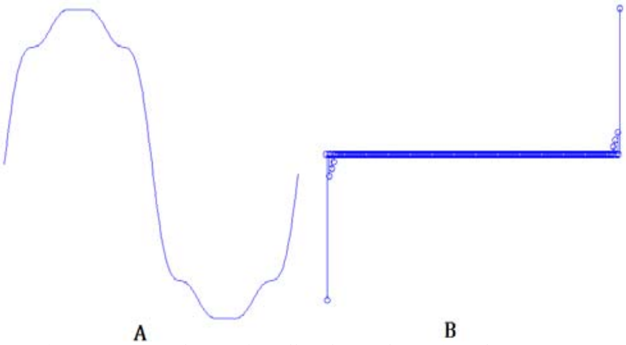

Figure 6. A) A harmonic polluted waveform B) Its frequency decomposition based on FFT

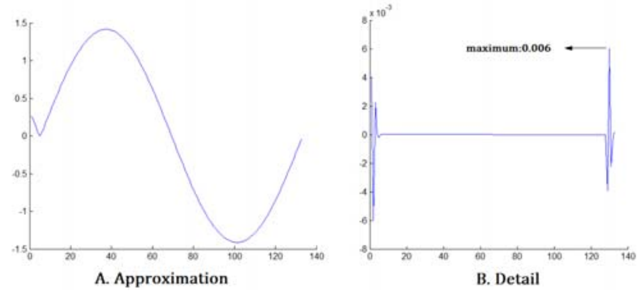

Figure 7. Decomposition of the waveform in Fig.6.A into approximation and details by DWT in one level based on db6 motherwavelet

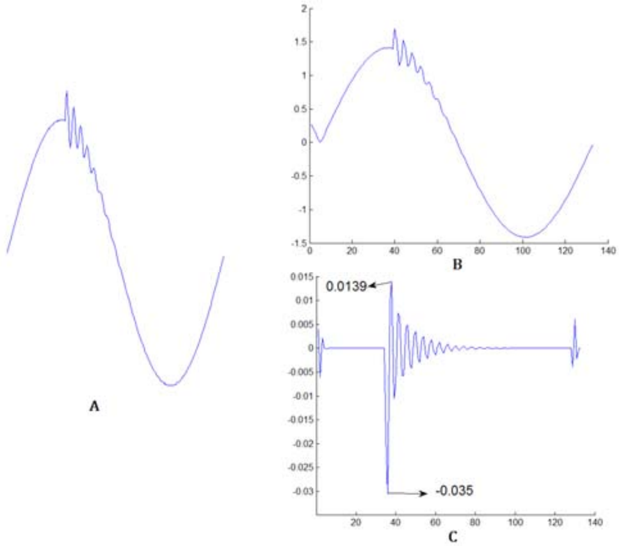

Figure 8. A) A sinusoidal waveform containing an oscillatory transient and its decomposition in one level by wavelet transformation into B) Approximation C) Detail based on db6 motherwavelet

V. Offline Processing of Power Quality Disturbances

In order to process the disturbances offline, voltage and current signals in each phase must be sampled and stored in a digital storage system, first. A digital data acquisition board with sampling frequency of at least 102.4 $^{\text{KHz}}$ and enough storage capacity based on the duration of analysis, according to table 2, is necessary for this purpose.

The stored data must be transferred to a digital processing system to be processed, then. Regarding to the large volume of data and complex computation algorithms, such a processing is out of capability of a personal computer due to the limited volume of RAM memory and the necessary duration of time for this purpose. Such a processing is possible only if a computer is specialized for online analysis of the disturbances. However, application of this system will not bring any advantages over a power quality analyzer. Therefore, approaches are investigated for offline processing of power quality disturbances.

VI. Big Data

Now-a-days, the amount of digital information is increasing at a high speed, due to the developments in digital processors and digital storage systems. Fig.9 shows a diagram of world's global information storage capacity evolution [15]. Therefore new fields of study are developed regarding to processing large datasets, known as *Big Data*. Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze and new architecture, techniques, algorithms, and analytics are required to manage it and extract hidden knowledge from it [16].

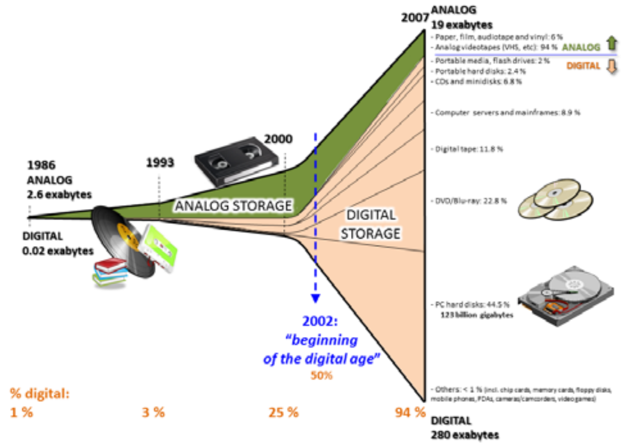
1986
ANALOG
2.6 exabytes
DIGITAL
0.02 exabytes
1993
2000
ANALOG STORAGE
DIGITAL STORAGE
2002:
"beginning of the digital age"
50%
2007
ANALOG
19 exabytes
- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 %
ANALOG
DIGITAL
- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %
- PC hard disks: 44.5 %
123 billion gigabytes
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)
% digital:
1 %
3 %
25 %
94 %
DIGITAL
280 exabytes

Figure 9. A diagram of world's global information storage capacity [15].

VII. Distributed Computing

Distributed computing is a field of computer science that studies distributed systems. A distributed system is a collection of individual computing devices that can communicate with each other [17]. A computer cluster is a class of distributed systems, which consists of a set of connected computers that work together so that in many respects they can be viewed as a single system [18].

The computer clustering approach usually connects a number of computing nodes, such as personal computers, via a fast local area network. The activities of the computing nodes are orchestrated by "clustering middleware", a software layer that sits atop the nodes and allows the users to treat the cluster as by and large one cohesive computing unit.

Cluster computing, provides an effective means for processing Big Data, since the storage and processing capabilities of a computer cluster is beyond the capabilities of a simple personal computer, however the set could be applied for a purpose like a PC. Special software packages are developed, as the clustering middleware, for processing of Big Data, such as Hadoop and MPI [19].

Figure 10. A simple computer cluster

VIII. Hadoop and Big Data

Hadoop is an open-source software written in Java, used for distributed processing of large datasets across large clusters of commodity servers [20].

The input data is divided into blocks of equal size, usually with size of 64MB, by a distributed file system named HDFS and stored on local disks of the machines in the cluster. Several copies of each block (typically 3 copies) is stored on different machines in order to increase reliability through replication. Fig.11 shows a diagram of the Hadoop Distributed File System (HDFS) [21].

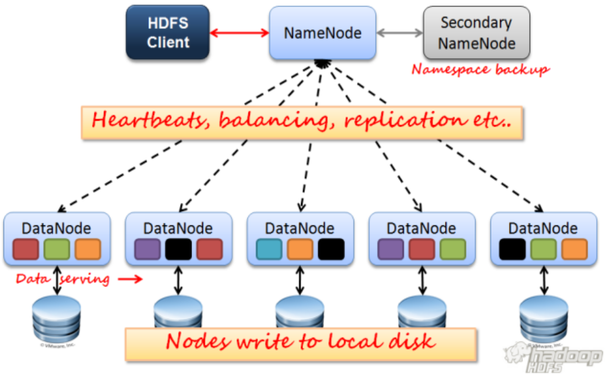

Figure 10. Basic Structure of Hadoop Distributed File System

Hadoop is based on simple programming model called MapReduce. MapReduce is a programming model and an associated implementation for processing and generating large datasets. 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. This process is depicted in Fig.11. The MapReduce master takes the location information of the input files into account and attempts to schedule a map task on a machine that contains a replica of the corresponding input data. Failing that, it attempts to schedule a map task near a replica of that task's input data [22].

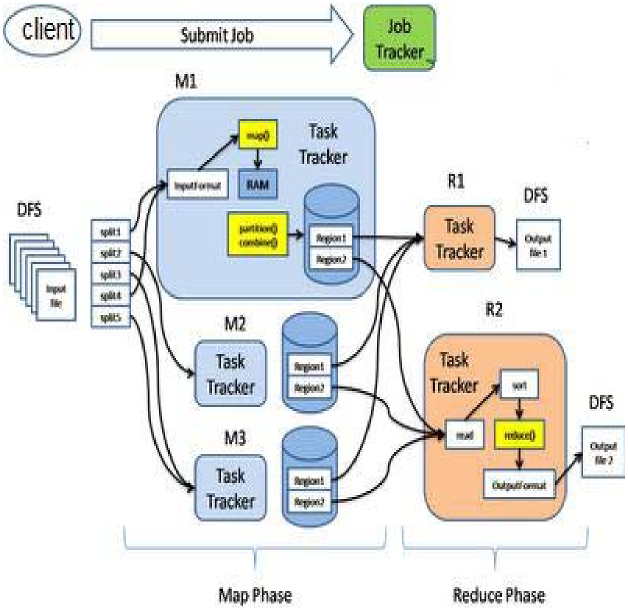

Figure 11. Operational Structure of Hadoop and MapReduce process.

IX. Application of Hadoop for Offline Processing of Power Quality Disturbances

Offline processing of power quality leads to a very big dataset, out of processing capability of a personal computer. This dataset is not so large for Hadoop and can be simply processed via Hadoop, in a network of computers. Such a network is present in most of electric distribution companies and even small industries for daily works and is out of use, outside working hours and during holidays. The application could also be performed in rent servers, such as *pay as you go cloud servers* [23].

Therefore, a data acquisition board is the only requirement for analysis of power quality disturbances in our proposed method and such a board does not cost more than $150 based on the price of the electronic chips and PCB included in this board. In comparison to a good power quality analyzer which costs at least $6000, this system will be much cheaper and application of this system will yield into a great deal of economic saving.

X. Experimental Results

A network of six computers was applied for analysis of power quality disturbances based on Hadoop. Datasets generated by Matlab were used as the data to be analyzed. Harmonics, oscillatory transients, notching, short duration variations and interharmonics were the five class of disturbances included in the generated datasets. DWT and FFT were used for detection of the disturbances, as mentioned in part 4. The execution time and number of maps, in this analysis is listed in Table 4.

TABLE IV. EXPERIMENTAL RESULTS OF RUNNING HADOOP ON A NETWORK OF SIX COMPUTERS

Data Volume (Gigabytes)	0.25	0.5	1	2	3	8	22
FFT's Execution time (s)	25	33	37	74	80	249	494
Number of maps for FFT	2	3	5	10	13	33	89
DWT's Execution time (s)	24	35	46	56	61	212	447
Number of maps for DWT	2	3	5	10	14	34	89
FFT & DWT's Execution time (s)	26	37	51	75	111	254	560
Number of maps for FFT & DWT	2	3	4	9	13	34	89

It can be concluded based on the results shown in Table 4, that analysis of a file with a volume of 22 ^{GB} based on FFT and DWT could be performed in less than 10 minutes on such a network. Therefore, Analysis of a file with volume of 98.88^{GB} , regarding to the samples saved during a day could be performed in less than 45 minutes, and analysis of a file with volume of 692.16^{GB} , regarding to the samples saved for a week could be performed in less than six hours, in such a computer cluster.

Increasing the number of computers in the cluster will yield into linear decrease in processing time and linear increase in processing capacity. The number of nodes in the computer cluster applied by some big data companies is listed in Table 5 [24].

TABLE V. THE NUMBER OF NODES IN THE COMPUTER CLUSTER APPLIED BY SOME BIG DATA COMPANIES

Company	YAHOO	LINKED IN	FACEBOOK	NETSEER
Nodes	42000	4100	1400	1050

XI. CONCLUSION

Rapid Developments of digital systems in digital age has led to large volume of digital storage capacities and significant processing capabilities of digital computers. Processing large datasets is a new field in computer science today and distributed processing is a solution for this purpose. Hadoop is an open-source software, developed for distributed processing of large datasets.

Processing a very large dataset is necessary for offline analysis of power quality disturbances, which is beyond the capabilities of a personal computer, but could be easily performed by Hadoop in a computer cluster. The network of computers present at electric distribution companies could be used as the computer cluster outside the working hours.

Application of Hadoop for offline analysis of power quality disturbances leads to a very cost effective system, resulting to a large amount of economic saving in comparison to the power quality analyzers and makes analysis of power quality disturbances possible for electric distribution companies with poor economic status.

REFERENCES

1) A. Kusko, M. Thompson, Power Quality in Electrical Systems, McGrawHill, 2007, pp 1-15.

2) IEEE Recommended Practice for Monitoring Electric Power Quality, IEEE Standard 1159, 1995 .

3) N. Mollayi and H. Mokhtari, "Classification of Wide Variety range of Power Quality Disturbances Based on Two Dimensional Wavelet Transformation", in *Proc. 1 st* Power Electronic & Drive Systems & Technologies Conference (PEDSTC), 2010, pp. 398-405.

4) DC. Dugan, MF. McGranaghan, Electrical Power Systems Quality, 2nd ed, McGrawHill, 2004, pp1-41.

5) EF. Fuchs, MAS. Masoum, Power Quality in Electrical Machines and Power Systems, Elsevier, 2008, pp1-44.

6) AV. Oppenheim, AS. Willsky, SH. Nawab, Signals and Systems, 2nd ed, Pearson International, 2014, pp 519-586.

7) B. Naodovic, "Influence of Instrument transformers on power system protection," Texas A&M University M.SC Thesis, 2005, pp 7-25.

8) AV. Oppenheim, RW. Schafer, JR. Buck, Discrete Time Signal Processing, Prentice Hall, 1998, pp 140-213, 541-669.

9) MHJ. Bollen, IYH. Gu, Signal Processing of Power Quality Disturbances, Wiley Interscience , 2006, pp 277-296.

10) SA. Deokar, LM. Waghmare, "Integrated DWT–FFT approach for detection and classification of power quality disturbances," International Journal of Electrical Power and Energy Systems, Elsevier, Vol. 61, Oct. 2014, pp 594-605.

11) G D. C. Robertson and O. I. Camps *et al.*, "Wavelets and electromagnetic power system transients," *IEEE Trans. Power Delivery*, vol. 11, Apr. 1996.

12) M. Karimi, H. Mokhtari, and M. R. Iravani, "Wavelet based on-line disturbance detection for power quality applications," *IEEE Trans. Power Delivery*, vol. 15, Oct. 2000.

13) S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.

14) I. Daubeshies, *Ten Lectures on Wavelets,* Philadelphia, Pennsylvania, SIAM, 1992, pp 1-102.

15) M. Hilber, P. Lopez, “The world’s technological capacity to store, communicate, and compute information,” Science, vol. 332, no.6025, Apr. 2011, pp. 60-65, doi:10.1126/science. 1200970.

16) D. Usha, A. Jenil APS, “A survey of Big Data processing in perspective of Hadoop and mapreduce, ” International Journal of Current Engineering and Technology, vol. 4, no. 2, Apr. 2014.

17) H. Attiya, J. Welch, Distributed computing: fundamentals, simulations and advanced topics. John Wiley & Sons, 2004, pp 1-14.

18) “Computer cluster” (Online). Available at: http://en.wikipedia.org/wiki/ Computer_cluster /. [Accessed: 2-June-2014].

19) J. Cohen, B. Dolan, et al , "MAD skills: new analysis practices for big data, " *Proceedings of the VLDB Endowment,* vol. 2, no. 2, Aug. 2009, pp 1481-1492.

20) “Welcome to ApacheTM Hadoop®!” [Online]. Available: https://Hadoop.apache.org/. [Accessed: 2-June-2014].

21) R. Chansler, H. Kuang, S. Radia, K. Shvachko, S. Srinivas, “The Hadoop distributed file system,”. In *Proc. IEEE Conf. Mass Storage Systems and Technologies (MSST), 2010,* pp. 1-10.

22) J. Dean, S. Ghemawat, “MapReduce: simplified data processing on large clusters,” *Communications of the ACM Magazine*, Vol.*51, Issue.*1, Jan 2008, pp. 107-113.

23) M. Armbrust, F. Armando, et al. "A view of cloud computing," *Communications of the ACM Magazine*, Vol.*53, Issue.*4, Apr 2010, pp. 50-58.

24) J. Wong, “Which Big Data Company has the World’s Biggest Hadoop Cluster?” (Online). Available at: http://www.hadoopwizard.com/which-big-data-company-has-the-worlds-biggest-hadoop-cluster /. [Accessed: 17-June-2014].