Application of a Hadoop-based Distributed System for Offline Processing of Power Quality Disturbances

Nader Mollaei[1], Seyyed Hadi Mousavi[2]
[1]Faculty of Computer and Industrial Engineering, Birjand University of Technology
[2]Faculty of New Sciences and Technologies, University of Tehran

Article Info

Article history:

Received Feb 1, 2017
Revised Apr 1, 2017
Accepted Apr 15, 2017

Keyword:

Apache Hadoop
Big data
Cloud computing
Distributed processing
Power quality
Power quality disturbances

ABSTRACT

Electric power quality is a critical issue for electric utilities and their customers and identification of the power quality disturbances is an important task in power system monitoring and protection. Offline processing of power quality disturbances provides an economic alternative for electric distribution companies, not capable of buying enough number of power quality analyzers for monitoring the disturbances online. Due to the wide frequency range of the disturbances which may happen in a power system, a high sampling rate is necessary for digital processing of the disturbances. Therefore, a large volume of data must be processed for this purpose for each node of an electric distribution network and such a processing has not yet been practical. However, thanks to the rapid developments of digital processors and computer networks, processing big databases is not so hard today. Apache Hadoop is an open-source software framework that allows for the distributed processing of large datasets using simple programming models. 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.

Corresponding Author:

Nader Mollaei,
Faculty of Computer and Industrial Engineering,
Birjand University of Technology,
Birjand, South Khorasan Province, Iran.
Email: mollayi@birjandut.ac.ir

1. INTRODUCTION

Electrical energy has been the most widespread type of energy used in the world for decades, due to its simpler transmission, distribution, conversion and usage and AC electrical power system has been in use for this purpose, worldwide. In this system, the best electrical supply is a constant magnitude and frequency sinusoidal voltage waveform. However, because of the non-zero impedance of the supply system, large variety of loads and 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.

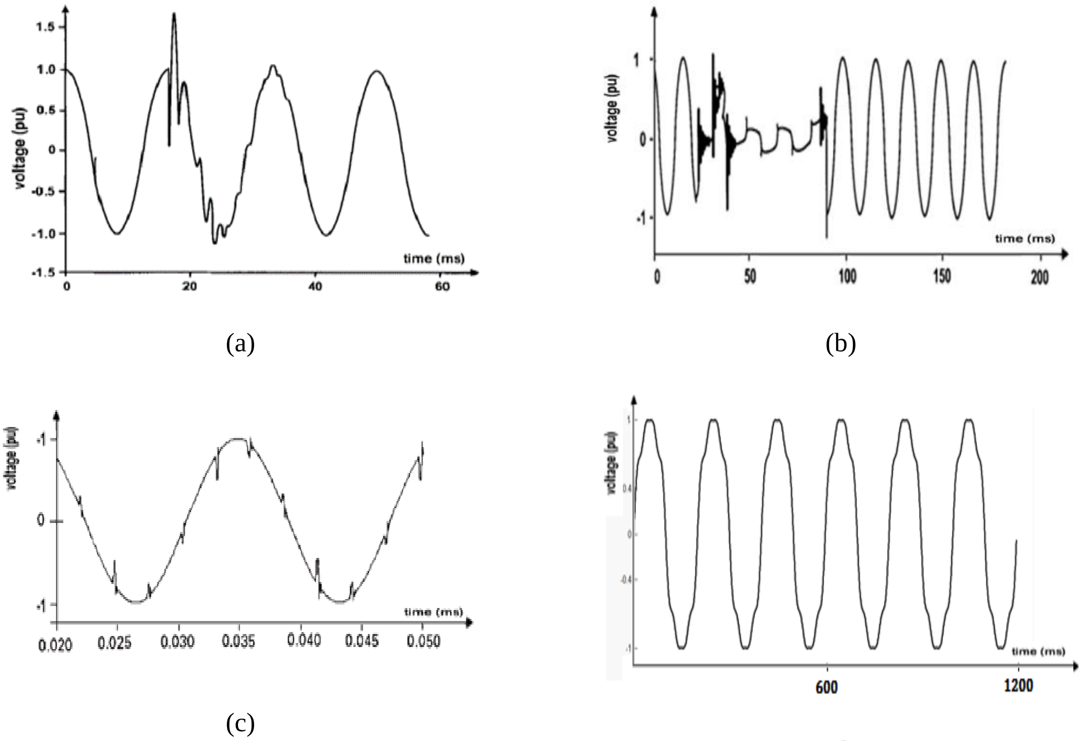
In recent years, there has been an increased concern for the quality of power due to the rapid developments of power electronic devices and their widespread use in industry. These devices are major sources of power quality problems, and on the other hand, they are much more sensitive to voltage disturbances than their counterparts in the past. Excessive heat, damage and service interruption in electric equipments, components aging and capacity decrease, malfunction of protection and measurement devices and reduction of power system efficiency are consequences of power quality disturbances [1]-[2]. 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 divided into seven classes of power quality disturbances based on IEEE 1159 standard [3].

To improve electric power quality, sources and causes of disturbances must be specified before taking any mitigation action. To achieve this purpose, events must be detected and classified [4]. 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 the power quality parameters is performed regularly in special points, but very often by electric distribution companies with poor economic status, such as the third 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.

Thanks to the developments in digital computers and computer networks, processing big databases is a new field of study in computer science, and special programming models and tools have been developed for this purpose. Apache Hadoop is an open-source software for processing large datasets with a simple programming model. 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 the possibility of widespread usage.

2. 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 any deviation from these conditions leads to a power quality disturbance. 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 [3]. Disturbances in each of the categories are characterized with a wide range of time and frequency domain characteristics. As a result, processing and characterization of the disturbances is a complicated task and is performed based on complex signal processing algorithms. Four different types of the disturbances are depicted in Figure 1 [5]. It can be obviously observed that the disturbances are completely different in their time and frequency domain characteristics.



(a)

(b)

(c)

(d)

Figure 1. Four different types of power quality disturbances: (a) Oscillatory transient caused by capacitor bank switching (b) Voltage sag caused by single line to ground fault (c) Notching caused by a three phase converter (d) Harmonic polluted waveform

Table 1. Categories and Typical Duration of Power Quality Disturbances, Defined by IEEE 1159 Standard

Category		Typical Duration
Transients	Oscillatory	$5^{\mu s} – 50^{ms}$ ●
	Impulsive	Shorter than 50 ^{ns} – longer than 1 ^{ms}
Short duration Variations (Sag, Swell, Interruption)		0.5 cycles – 1^{min}
Long duration Variations		Longer than 1 min
Voltage Imbalance		Steady State
Waveform Distortion		Steady State
Voltage Fluctuations (Frequency Range < 25 Hz)		Intermittent
Power Frequency Variations		$<10^{s}$

3. 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]. Neglecting high frequency voltage transients, with duration of less than 20µs, or frequency contents of higher than 50kHz, a minimum sampling frequency of 100 kHz is necessary for this purpose. A higher sampling frequency may not lead to 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. A sixteen bit sampling resolution is also necessary for an acceptable precision, resulting in quantization error of less than 10 mv. Therefore, the number of cycles and the data volume necessary for processing each of the voltage or current waveforms in various time intervals is calculated as listed in the first two rows of Table 2. The total recording volume must be six times larger, since the voltage and current in three phases must be processed in a three phase system for each node of an electric power distribution system [8].

Table 2. Data Volume Required for Recording Power Quality Disturbances with 2048 Samples/Cycle in a 50^{Hz} System

Duration	One Second	One Hour	One Day	One Week
Number of cycles in each signal	50	180000	$4.32 * 10^{6}$	$30.24 * 10^{6}$
Data Volume for Each Signal	200^{Kb}	703.13^{Mb}	16.48^{Gb}	115.36^{Gb}
Total Data Volume	1200^{Kb}	4.12^{Gb}	98.88^{Gb}	692.16^{Gb}

In order to process the recorded data in a digital computer, each sample must be converted to the x86 floating point format with length of eight bytes. Therefore, the volume of the dataset which must be processed for 100 nodes of an electric distribution network will be 400 times larger than the volume of the recorded data in each point, as listed in Table 3.

Table 3. Total Data Volume Required for Processing Power Quality Disturbances in 100 Nodes of an Electric Power Distribution Network

Duration	One Second	One Hour	One Day	One Week
Data Volume	468.75^{Mb}	1.61^{Tb}	38.625^{Tb}	270.5^{Tb}

4. 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 have been developed for this purpose. Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) have traditionally been used for detection of steady state disturbances and transients, respectively [9].

Discrete Fourier Transform decomposes a signal into its frequency components and is implemented by Fast Fourier Transform (FFT) algorithm. Discrete Wavelet Transform decomposes a signal into approximation and details which contain the low and high frequency components of the signal, respectively. Steady state disturbances can be identified via the frequency components calculated by DFT and the details resulting from DWT are suitable for characterization of transients. The calculation and frequency characteristics of wavelet decomposition mainly depend on the selected motherwavelet [10]–[13].

Application of the Biorthogonal 3.1 mother wavelet has been proved to be efficient for classification of power quality disturbances [14]. A pure sinusoidal waveform, a harmonic polluted sinusoidal waveform and a sinusoidal waveform containing a transient are depicted in Figures 2a-4a. The amplitude of DFT and the detail of DWT in one level calculated for them are shown in Figures 2b,c -4b,c.

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

1. Frequency components with non-negligible amplitude in comparison with the amplitude of the main frequency component.
2. The total energy of the detail based on the Biorthogonal 3.1 motherwavelet.

5. 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. Therefore, a digital data acquisition board with sampling frequency of at least 102.4 KHz and enough storage capacity based on the time duration of analysis, according to Table 2, is necessary for this purpose.

The stored data must then be transferred to a digital processing system. Regarding to the large volume of data and complex computational algorithms, such a processing is beyond the capability of a personal computer due to the limited volume of memory and the necessary processing time for this purpose. Therefore, approaches are investigated for offline processing of power quality disturbances. A sinusoidal waveform with an oscillatory transient, the amplitude of its DFT, the detail of its DWT as shown in Figure 4.

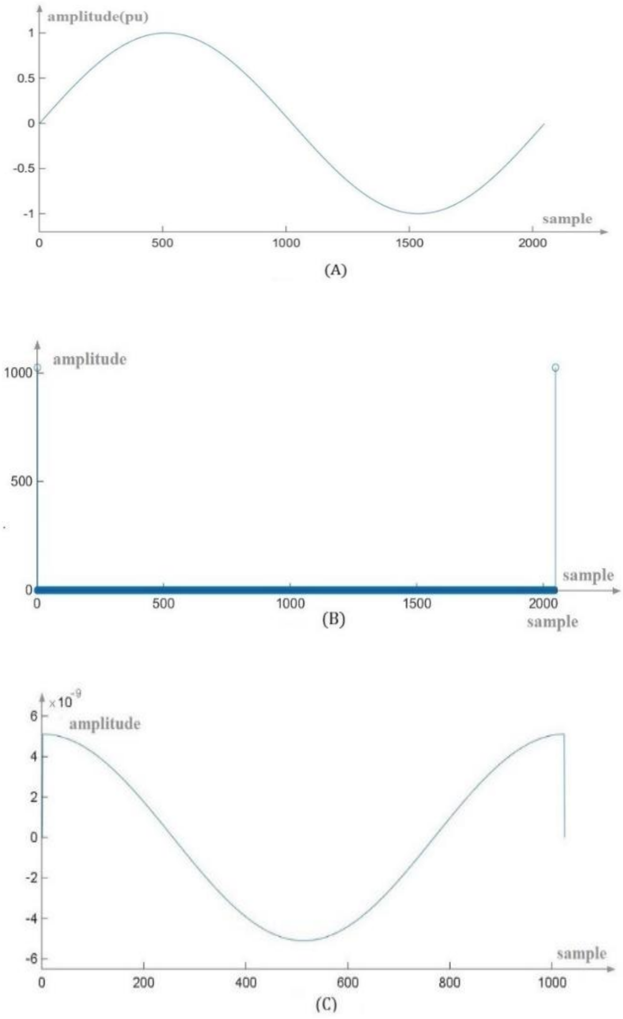


Figure 2. (A) A pure sinusoidal waveform (B) The amplitude of its DFT (C) The detail of its DWT

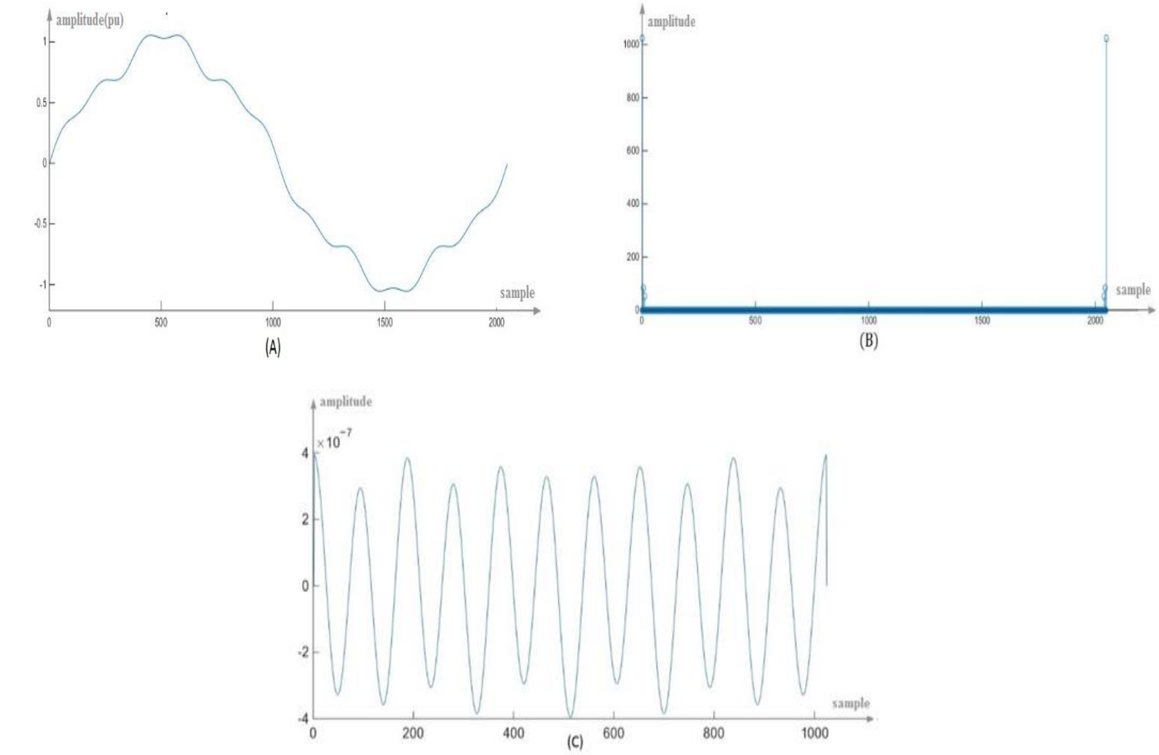


Figure 3. (A) A harmonic polluted sinusoidal waveform (B) The amplitude of its DFT (C) The detail of its DWT

6. 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. Figure 5 shows a diagram of world's global information storage capacity evolution from 1986 to 2007 [15]. Therefore, new fields of study have been developed for 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 and extract hidden knowledge from it [16].

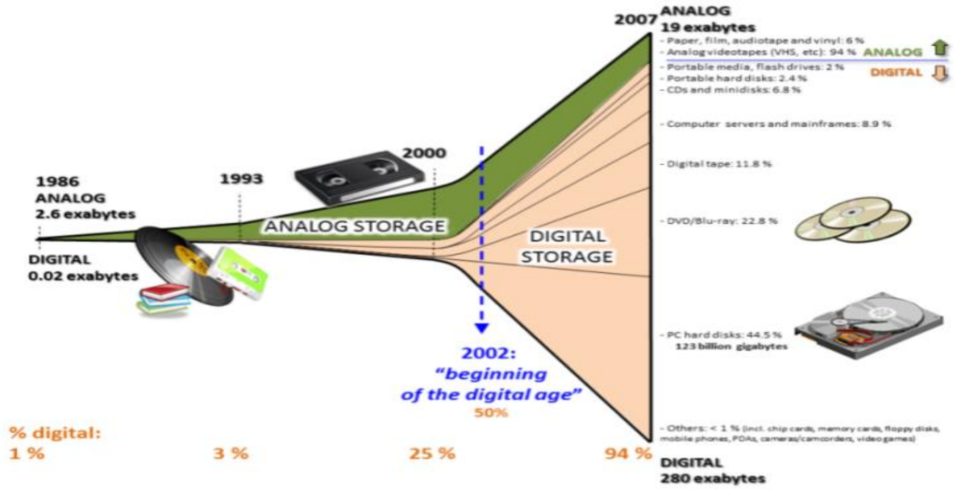


Figure 5. A diagram of world's global information storage capacity from 1986 to 2007 [15]

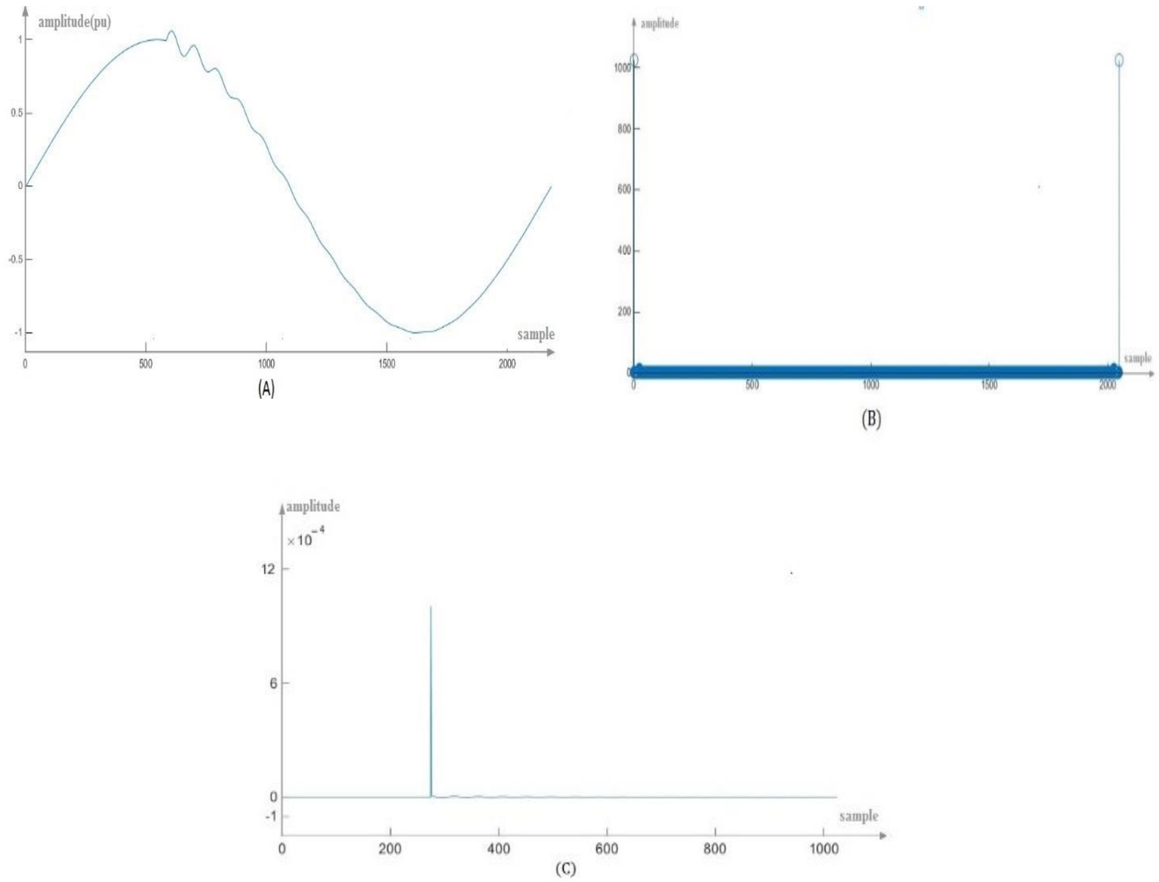


Figure 4. (A) A sinusoidal waveform with an oscillatory transient (B) The amplitude of its DFT (C) The detail of its DWT

7. 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, consisting of a set of connected computers that work together so that in many aspects 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 a large 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 can be applied for a purpose like a PC. Special software packages have been developed, as the clustering middleware, for processing Big Data, such as Hadoop and Spark [19].

8. 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. In a Hadoop-based distributed system, the input data is divided into blocks of equal size, usually with a 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) are stored on different machines in order to increase reliability through replication. Figure 6 depicts the basic structure of Hadoop Distributed File System (HDFS) [20].

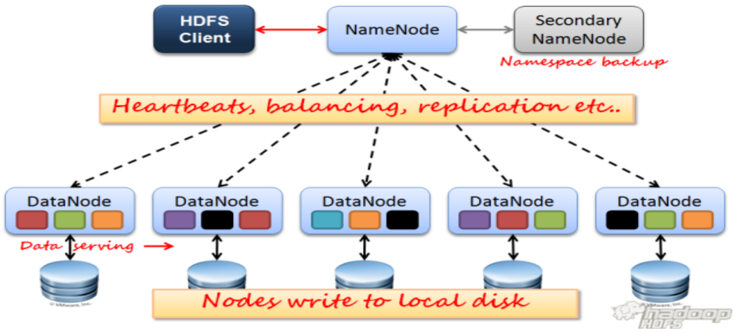


Figure 6. Basic Structure of Hadoop Distributed File System

Hadoop is based on a 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 [21]. This process is depicted in Figure 7. 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].

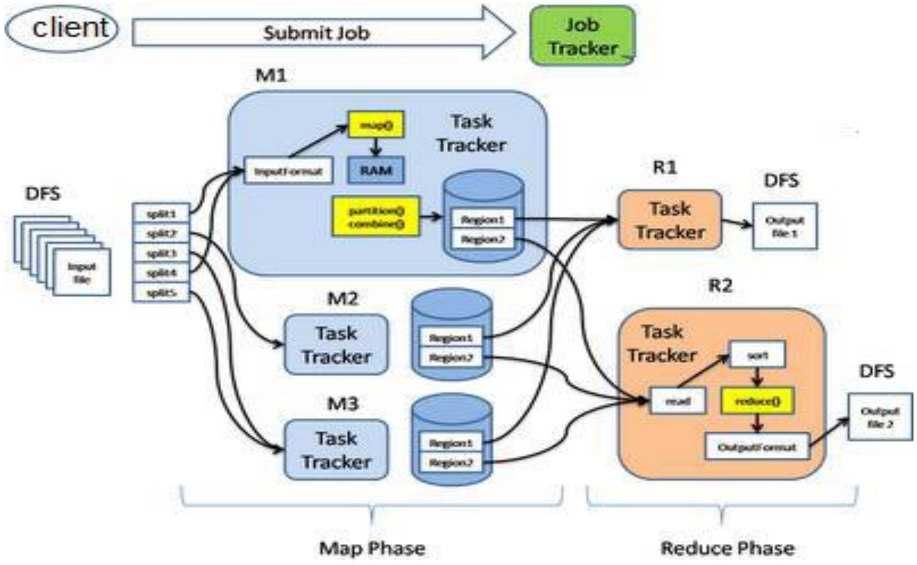


Figure 7. Operational Structure of Hadoop and MapReduce process

8.1. Using Hive UDFs to Implement Parallel Processing

Apache Hiveis a data warehouse software that facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. Processing big data in the Hadoop is mainly done by coding the map and reducing the jobs (e.g. in java). This way needs the map/reduce programming skill. But we make use of Hive and Hive UDFs so that we just need to write main processing code in java that needs to be applied on each sample. The Hive engine will use the Hadoop infrastructure to distribute the jobs through the clusters nodes and process all samples parallel. Using Hive UDF [24] and Hive query the processing is separated from the map/reduce model, so that parallel processing can be done without writhing map/reduce jobs. The other benefit of writing UDFs is that we can run our job from the command line and any other Hive interfaces (such as Hive command line and Hive java API).

9. APPLICATION OF HADOOP FOR OFFLINE PROCESSING OF POWER QUALITY DISTURBANCES

Offline processing of power quality disturbances leads to a very big dataset, out of the processing capability of a personal computer. This dataset is not so large for a Hadoop-based system 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 the holidays. The application can 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 costs a lot cheaper than a power quality analyzer. In conclusion, application of this system in a large scale will yield into a great deal of economic saving.

10. EXPERIMENTAL RESULTS

The implementation was built on a Hadoop cluster with ten virtual machines. Each cluster had the following configurations. All the ten virtual machines were hosted in two HP servers. A large database of power quality disturbances and pure sinusoidal waveforms was provided for execution of the experiments. The database included harmonics, oscillatory transients, notching, short duration variations and interharmonics, generated by Matlab command line instructions based on their standard characteristics. DWT and DFT were used for detection of the disturbances, as mentioned in part 4. The execution time and number of maps in this analysis are listed in Table 5.

Table 4. Cluster Machines Configurations

Machine	CPU	Ram
Cluster Nodes (virtual machine)	16 *1.995 GHz (shared with all VMS)	4GB
ESX Server (HP DL380 G8)	16 *1.995 GHz	64GB

Table 5. Experimental Results of Running Hadoop on ten Virtual Machines Hosted in Two HP Servers

Processed Million Cycles	0.02	0.1	0.2	0.3	0.5	1	1.5	2	3
Number of generated maps	4	11	19	26	43	83	166	205	245
Mean Processing Time (s)	54.5	109.3	214	330.5	591	1199	2179	2854	4316

As depicted in Figure 8, the processing time increases almost linearly with the increase of cycles to be processed, after 1.5 million cycles. The different behavior of this system before this point is due to the fact that Hadoop is designed for processing large databases.

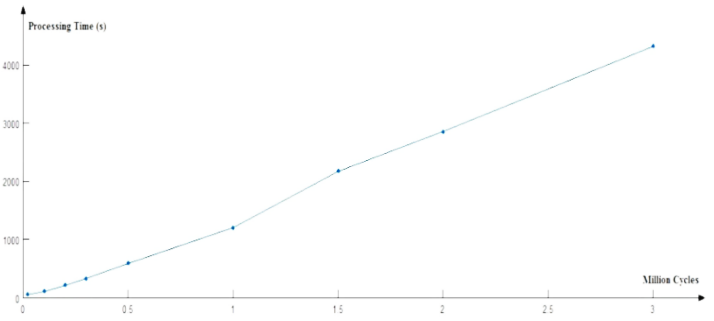


Figure 8. The processing time based on the million cycles of power system signal in the proposed system

As a result, the three phase voltage and current signal recorded for one day, consisting of 25.92 million cycles, can be processed in less than 10.5 hours in this system. This task can be completely performed by the server and the computer network, at an electric distribution company, outside the working hours. If the weekends are also taken into consideration, at least two points of a power system can be analyzed based on Hadoop, outside the working hours, in such a system each week. Therefore, power quality analysis of more than 730 points can be performed in this system, each year, which is a great advantage over the traditional power quality analyzers.

The data can be transferred to the server via the empty bandwidth of the power line communications system tools, the fiber optic cables used for primary power system communications, or the internet network, where possible. It is also possible to store the data and then transfer it to the server to be processed. Increasing the number of computers in the cluster will yield into linear decrease in processing time and linear increase in processing capacity.

11. CONCLUSION

Rapid Developments of digital systems in digital age, has led to large volume of digital storage capacities and significant processing capability 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 server and the computer network, present at electric distribution companies can 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 with a great deal of economic saving in case of wide-spread application in comparison to the power quality analyzers and enables electric distribution companies with poor economic status to monitor power quality disturbances.

REFERENCES

[1] A. Kusko and MT. Thompson, “Power quality in electrical systems,” *McGraw-Hill,* pp. 1–14, 2007.

[2] M.J. Ghorbani and H. Mokhtari, “Impact of Harmonics on Power Quality and Losses in Power Distribution Systems,” *International Journal of Electrical and Computer Engineering (IJECE),* vol/issue: 5(1), p. 166, 2015.

[3] J.C. Smith, et al., "IEEE Recommended Practice for Monitoring Electric Power Quality," *IEEE Std,* p. 1159, 1995.

[4] N. Mollayi and H. Mokhtari, “Classification of wide variety range of power quality disturbances based on two dimensional wavelet transformation,” in *Proceeding of 1st Power Electronic & Drive Systems & Technologies Conference (PEDSTC),* pp. 398–405, 2010.

[5] E. Fuchs and M. A. Masoum, “Power quality in power systems and electrical machines,” *Academic press,* pp. 1-44, 2011.

[6] A. V. Oppenheim, et al., “Signals and systems,” *2nd ed. New Jersey: Prentice Hall,* pp. 519-586, 1997.

[7] B. Naodovic, “Influence of instrument transformers on power system protection,” *Doctoral dissertation. Texas A&M University, pp. 7-25, 2005.*

[8] A. V. Oppenheim, et al., “Discrete-time signal processing,” *Prentice-hall Englewood Cliffs, vol. 2, pp. 140-213, 541-669, 1989.*

[9] M. H. Bollen and I. Gu, “Signal processing of power quality disturbances,” *John Wiley & Sons, vol. 30, pp 277-296, 2006.*

[10] S. A. Deokar and L. M. Waghmare, “Integrated DWT–FFT approach for detection and classification of power quality disturbances,” *Int. J. Electr. Power Energy Syst*., vol. 61, pp. 594–605, 2014.

[11] D. C. Robertson, et al., “Wavelets and electromagnetic power system transients,” *Power Deliv. IEEE Trans. On, vol/issue: 11(2), pp. 1050–1058, 1996.*

[12] M. Karimi, et al., “Wavelet based on-line disturbance detection for power quality applications,” *Power Deliv. IEEE Trans. On, vol/issue: 15(4), pp. 1212–1220, 2000.*

[13] I. Daubechies, “Ten lectures on wavelets,” *SIAM, vol. 61, pp 1-102, 1992.*

[14] W. A. Adil, et al., “Investigation of suitable Mother Wavelet Transform Functions for Detection of Power System Transient Disturbances,” in *Proceeding of First International Conference on Modern Communication & Computing Technologies,* Nawabshah, Pakistan, 2014.

[15] M. Hilbert and P. López, “The world’s technological capacity to store, communicate, and compute information,” *science, vol/issue: 332(6025), pp. 60–65, 2011.*

[16] D. Usha and A.P.S. Aslin Jenil, “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 and J. Welch, “Distributed computing: fundamentals, simulations, and advanced topics,” *John Wiley & Sons, vol. 19, pp. 1-14, 2004.*

[18] R. Buyya, “High performance cluster computing,” *N. J. Frentice, pp. 3-48, 1999.*

[19] J. Cohen, et al., “MAD skills: new analysis practices for big data,” *Proc. VLDB Endow*., vol/issue: 2(2), pp. 1481–1492, 2009.

[20] K. Shvachko, et al., “The hadoop distributed file system,” in *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST),* pp. 1–10, 2010.

[21] R.P. Padhy, “Big data processing with Hadoop-MapReduce in cloud systems,” *International Journal of Cloud Computing and Services Science*, vol/issue: 2(1), p.16, 2013.
[22] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” *Communications of the ACM, vol/issue: 52(1), pp. 107–113, 2008.*
[23] M. Armbrust, et al., “A view of cloud computing,” *Communications of the ACM, vol/issue: 53(4), pp. 50–58, 2010.*
[24] A. Thusoo, et al., “Hive: a warehousing solution over a map-reduce framework,” in *Proceeding of VLDB Endow*., vol/issue: 2(2), pp. 1626–1629, 2009.

BIOGRAPHIES OF AUTHORS

Nader Mollaei was born in Mashhad, Iran in 1985. He received his B.Sc. degree in electrical engineering from Shahrood University of Technology, Shahrood, Iran and his M.Sc. degree in power electronics from Sharif University of Technology, Tehran, Iran in 2007 and 2010, respectively. Since September 2010, he has been a lecturer in Birjand University of Technology, Birjand, Iran. His research interests include power electronics, power quality, digital signal processing, pattern classification and processing of power quality disturbances.

Seyyed Hadi Mousavi was born in Birjand, Iran, in 1987. He received the B.E. degree in Information Technology engineering from the University of Birjand, Birjand, Iran, in 2009, and the Master degree in Information Technology-Ecommerce from the Amirkabir University of Technology (Tehran Polytecnic) Tehran, Iran, 2011 and he started his Ph.D. in 2013. In 2010, he joined the Cloud Computing Research Center (CRC), Amirkabir University of Technology, as a Researcher, and since then he started his research on Big Data and Cloud Computing solutions. Since October 2012, he has joined the Department of Computer and Information Technology Engineering, Birjand University of Technology, as a lecturer. In 2013 he was awarded best researcher in Birjand University of Technology. His current research interests include Cloud Computing, Big Data Application, Hadoop Solutions, Network Security and Machine learning on Big Data.