Hadoop-based framework for big data analysis of synchronised harmonics in active distribution network

Zijian Cao1, Jin Lin1, Can Wan2, Yonghua Song1,2, Gareth Taylor3, Maozhen Li4
1Department of Electrical Engineering, Tsinghua University, Beijing, People’s Republic of China
2College of Electrical Engineering, Zhejiang University, Hangzhou, People’s Republic of China
3Brunel Institute of Power Systems, Brunel University, London, UK
4School of Engineering and Design, Brunel University, London, UK
E-mail: canwan@zju.edu.cn

Abstract: Synchronised harmonic analysis that utilises GPS signals to synchronise harmonic measurements enables advanced power quality analysis in smart grid. However, it is difficult to put this promising technology into practice due to an extremely high requirement on data storage and processing. To overcome this problem, this study presents a novel Hadoop-based framework for big data analysis of synchronised harmonics in distribution networks. The proposed framework facilitates harmonic data processing associated with data storage and advanced analysis based on big data techniques. Under this framework, a big matrix multiplication algorithm based on MapReduce programming model solving the harmonic state estimation problem is deployed in the stage of data processing. A MapReduce-based harmonics distortion calculation is implemented as an advanced algorithm and verify the effectiveness of the proposed framework.

1.1 Introduction

With the development of smart grid, large scale of intelligent sensing, monitoring and measurement systems are established in smart grid to reach deeper situation awareness, prompter responses and safer power supply in an automatic way with less human involvement [1, 2]. A variety of advanced technologies including phasor measurement units (PMUs) [3], smart meters [4], and advanced metering infrastructure [5] have been utilised to produce more advanced sensing network. The majorities of these technologies are data-intensive. A huge amount of data will be generated, transmitted, stored and analysed after the deployments of sensing systems. For instance, based on the Navigant Research Report, it is estimated that the smart meters installed worldwide will surpass 1.1 billion by 2022 [6], which is expected to bring more than 2 petabytes of data annually [7]. Moreover, the data requirement may arise much more rapidly with the maturing of Internet of things [8] and Energy Internet [9].

Therefore, the concept of ‘Big Data’ which is initially developed from the internet and information industry is becoming widely adopted in smart grid. Several common characteristics of big data technology are described as follows [10, 11]:

i. The volume of dataset can be extremely large, which is usually beyond the capability of traditional systems for data storage, management and analysis.
ii. More valuable intelligences hidden beneath the raw data can be exposed by big data technologies.

Lots of efforts have been paid to study big data techniques and platforms. Big data techniques cover the acquisition, conditioning, and evaluation of big data, and big data platforms include software and hardware that support the deployment of the big data techniques across clusters of computer units [11]. Under big data environment, parallel algorithms and distributed approaches are proposed for data mining among multi-database by database classification and selection [12]. As a powerful tool to discover knowledge and make decision, machine learning algorithms can be scaled up to cope with the large dataset [13, 14]. Except for the big data techniques, some novel platforms suitable for big data are also developed recently. Hadoop is one of the most famous big data processing platforms [15], including Hadoop distributed file system (HDFS), MapReduce programming model [16] and so on. In addition, Hadoop [17] is modified from Hadoop to better support iteration applications, and Spark is proposed to support cyclic data flow and in-memory computing [18]. Some preliminary studies have been conducted to employ big data technologies to smart grid. PMU data are stored and analysed by a Hadoop-based platform which is also used to study the frequency incidents of National Grid in United Kingdom [19]. A cloud-based platform is proposed in [20], which offers an adaptive information integration pipeline and scalable machine learning models for demand forecasting to support big data analysis in smart grid. A cost-oriented model in order to optimise the allocation of cloud computing resources for a cloud-based information communication technology (ICT) infrastructure in demand side management is proposed in [21]. The technical challenges and risks for applying big data techniques in distribution networks are discussed in [7]. Most existing studies on applying big data technologies in power systems mainly focus on dealing historical data, not closely relevant to the operational-level issues in power systems.

Synchronised harmonics that applies GPS to synchronise harmonic monitoring of power systems has been developed since last decade [22, 23]. Although the high cost has limited the utilisation of synchronised measurements at distribution level at first, but recent achievements across electronics sector have driven the cost of PMUs dropped dramatically in price. Moreover, it has reached a truing point that has made the PMU an attractive tool across the utility environment including distribution system and embedded generation [24] but the technology is difficult to be used in practice mainly due to the extremely high requirements on storage and processing capabilities. However, due to the deregulation of energy markets, the electricity price can depend on the harmonics quality [25]. Especially in the emerging active distribution network, more and more electronic-based devices, such as distributed generators, electrical vehicles and storages, are connected into the low-voltage networks via inverters, which will lead to more severe harmonics issues [26]. Meanwhile, the characteristics of the propagation and interaction among multi-
sources make the harmonics issues no longer single-point but network-level [27]. So it is an urgent need to monitor and analyse the harmonic situations and provide corresponding energy service based on the harmonic analysis results. Similar with the state estimation analysis for fundamental wave, the harmonics state estimation has been brought up for decades as an effective way to analyse harmonics situations. The classic harmonics state estimation (HSE) algorithm is inherited from state estimation for fundamental frequency based on the weighted least squares method [28]. Time-domain HSE based on Kalman filter method is proposed in [29, 30]. The vital challenge to applying synchronised harmonics analysis in distribution network is how to process such huge data efficiently to satisfy the practical requirements. Therefore, it is necessary to develop a platform capable of dealing the huge harmonic data.

This paper proposes a Hadoop-based framework for big data analysis of synchronised harmonics in distribution networks. Based on this platform, MapReduce-based harmonic state estimation (MHSE) algorithm and total harmonics distortion (THD) calculation algorithm are implemented. The comprehensive framework gives a complete harmonic data processing procedure consisting of data storage, data pre-processing and advanced analysis based on big data techniques. The proposed big data framework can provide an alternative to solve this issue and make the promising technology more practical for advanced power quality analysis, such as power flow and state estimation analysis. Numerical experiments are carried out on this platform to demonstrate the efficiency and the accuracy of this framework for MHSE and THD calculation applications. The primary contributions of this work are listed as follows:

i. A Hadoop-based synchronised harmonics analysis platform is proposed in this paper, which has practical potential on both on-line operational applications and off-line historical data processing applications.

ii. HDFS-based data storage architecture and MapReduce-based HSE (MHSE) algorithm are deployed on the platform to store and process huge synchronised harmonics data efficiently.

iii. Basic case and an extended case with thousands of nodes are studied to evaluate the proposed platform and algorithm from the perspectives of efficiency and accuracy.

The remaining paper is organised as follows: Section 2 introduces the motivation and the basic idea of the application of the Hadoop technology for the data storage and processing of synchronised harmonics. The MapReduce algorithm of state estimation on synchronised harmonics and the implementation on a Hadoop-based framework is proposed in Section 3. Comprehensive numerical case studies are carried out in Section 4 to demonstrate the effectiveness of the proposed framework. The conclusion of this paper is presented in Section 5.

2 Big data issue of synchronised harmonics

2.1 Preliminaries on synchronised harmonics

In traditional distribution networks, the influence of harmonics is limited due to the relative low harmonics level. However, in the booming active distribution network that contains considerable amount of ‘active equipment’ such as distributed generations [31, 32], electrical vehicles, storages and other power electronic interfaced devices, power quality issues, especially harmonic issues, have become critical for the efficient and secure operation of distribution networks. Those inverter-connected non-linear devices will generate much more harmonics. More worse, the harmonics can propagate and interact between the large number of devices connected to the low-voltage network through inverters [27], which may lead to severely harmful consequences such as high resonance voltages or current [26, 33]. Both simulation results and field measurements have shown that harmonics violations would occur in a distribution network with high penetration of PV systems even each PV system satisfies the harmonics emission standard individually [26, 28]. The harmonics generated by non-linear components in the distribution network will significantly reduce the lifetime of these ‘active equipment’ and even prevent the interconnection of these ‘active equipment’ due to the violation of power quality requirement [27]. This has become a major concern for the distribution utility.

Harmonics have become the bottleneck for the development of active distribution network. In the active distribution network, harmonics are injected from multi-sources and propagate through the whole network. Then the harmonics issue becomes a network-level problem [28]. Especially in the deregulated utility network, each customer involved will focus on its own benefits, so it is vital to be aware of the harmonics distributions and locate the sources of harmonics pollution [30]. Harmonics state estimation is a powerful tool to accomplish source location and distribution awareness of harmonics. A number of techniques have been proposed for the awareness of harmonic in modern distribution networks [29, 34], most of which strongly rely on the calculation of harmonic power flow and state estimation. However, accurate calculations on harmonic flows cannot be implemented as conveniently as normal power flow, due to that high-order harmonic measurements should be synchronised first to ensure that collected data reflect the same snapshot of the network. To overcome these difficulties, synchronised harmonics is proposed [22, 23], the concept of which can be illustrated in Fig. 1. The timestamps of GPS signals are attached to every measurement record so that high-order harmonic monitoring can be synchronised, and then the harmonics flow and state estimation can be calculated. The practical synchronised harmonics measurement system has been deployed over a decade using phasor measurement technology in New York State [35]. Harmonics data synchronised by GPS clock is also used for harmonics state estimation in Japan [36]. Especially in active distribution network, PMU-based synchronised monitoring system is deployed to monitor more accurate dynamic behaviour of the electric network [37, 38]. For harmonics issues, GPS-based synchronised harmonics measurement is also proposed for distribution network [23] and further step to establish such a synchronised measurement system with less cost based on PMU has been studied in [39]. With the gradual development of PMU devices, it will become easy and cost-friendly to establish a synchronised harmonics measurement system in distribution networks.

The main difficulty to construct such a network is the extremely high requirements of storage and processing capabilities. A large amount of raw data can be generated based on GPS synchronised power quality monitoring system composed of a great number of harmonic monitoring devices [22, 40]. Therefore, a powerful data processing platform is necessary for synchronised harmonic applications, which is presented in the next subsection.

Fig. 1 GPS-based synchronised harmonics measurement system

[This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/) IET Gener. Transm. Distrib., 2017, Vol. 11 Iss. 16, pp. 3930-3937 3931]
2.2 Data storage of synchronised harmonics

The data-intensive feature of synchronised harmonic would introduce significant challenges to data storage, processing and statistical analysis. The volume of raw data recorded by an individual harmonic monitoring instrument has been considerably large. In practice, the number of data points recorded by an individual instrument is listed in Table 1. If each data point is stored as a 32 bit float number, the recorded data volume per day can be calculated as following:

\[32 \text{ bit} \times 4,320,000 \times 4 \approx 69.12 \text{ MB}\]

Data volume will increase more significantly if a synchronised monitoring system is taken into account which uses GPS timestamps to synchronise measurement snapshots [22, 40]. According to the latest development, the largest power quality monitoring system in China is established in the distribution network of Guangzhou [41], involving more than 1000 harmonic monitoring instruments in the medium-voltage (MV) distribution network (110/10 kV) and about 46,000 instruments in the low-voltage (LV) network (380 V). In Guangzhou's distribution network, the scale of a historical harmonic database containing one-year raw data recorded by 1000 monitoring instruments in MV network can reach over 25TB. In practical applications, it might not be necessary to process one-year data at one time. Customers could choose the time scope and resolution based on their specific requirements. For some off-line historical data analysis applications, monthly, seasonal or even annually data could be extracted from the whole historical data set and then processed. For on-line application scenario, maybe only minute-level data set need to be processed at one time.

2.3 Data processing of synchronised harmonics: harmonic state estimation algorithm

As data quality of monitoring instruments could be influenced by complex environments of distribution networks, raw data cannot be directly used for harmonic analysis. Furthermore, monitoring instruments are only installed on a limited number of nodes due to the limitation of investments. Harmonic state estimation is regarded as a common way to clear bad data and achieve full awareness of the entire network [28, 34, 36, 42]. One classic algorithm based on weighted least squares method gives the form of results [36], expressed as

\[X^* = (H^*WH)^+H^*WZ_m\]  \hspace{1cm} (1)

where \(Z_m\) represents the measurement matrix, \(X^*\) is the estimation matrix, \(W\) is the weighting matrix and \(H\) represents Jacobian matrix. In most cases, the Jacobian matrix \(H\) and weighting matrix \(W\) are constant for HSE. Consequently, the estimation matrix can be transformed as

\[X^* = AZ_m\]  \hspace{1cm} (2)

where the matrix \(A\) is constant. Accordingly, the calculation of HSE is converted to a matrix multiplication. The raw data is contained in the measurement matrix \(Z_m\) and each column of matrix \(Z_m\) represents a synchronised snapshot of harmonic measurement. In practise, the dimension of matrix \(Z_m\) can be extremely large. As mentioned above, in a system with 1000 monitoring instruments, the dimension of matrix \(Z_m\) containing one year measurement data is as large as 4000 \times 10,512,000. Such a big matrix multiplication is extremely time-consuming and sometimes infeasible under current computing environment of power systems. Furthermore, in most cases of distribution networks, the dimension of the estimation matrix \(X^*\) is even larger than the measurement matrix \(Z_m\). It is obvious that the statistical index calculation on the big matrix \(X^*\) also requires advanced big data technologies.

Observability analysis is another important question for any kind of state estimation. Observability analysis could be basically classified into two types: numerical analysis and topological analysis. An observability analysis method for HSE using classical topological analysis with the same computational requirements as for the conventional SE is proposed in [28] under two practical conditions. To overcome the limited measurements for HSE in distribution network, an estimation algorithm based on singular value decomposition is proposed which is capable of estimating the variables in observable islands while the rest state variables remaining unknown [43]. Actually, different HSE algorithms have different requirements for the observability of distribution network. And in practical, some pseudo-measurements rescheduled from a priori knowledge are commonly used in state estimation (SE) in distribution network [44]. The optimal deployment strategy of the measurement device is another study area [45] which is out of the scope of this paper. For simplicity, the measurement deployment problem is not considered in this paper. In this paper, it is assumed all the harmonics data calculated during the simulation are used in the HSE as direct measurements or pseudo-measurements. Under this condition, the observability can be fully guaranteed and the data size can be big enough to test the ability of the proposed platform.

3 Hadoop-based synchronised harmonics analysis: framework and MapReduce algorithm

To overcome the challenges of big data storage and processing, a Hadoop-based technology for synchronised harmonic analysis is proposed in this section. A comprehensive framework is proposed based on Hadoop structure; then a multiplication algorithm for big matrix is proposed for HSE based on MapReduce programming.

3.1 Hadoop-based framework for synchronised harmonics analysis

Hadoop is an open source software structure that enables the distributed storage and processing of large data sets across clusters of commodity hardware using simple programming models [15]. The two core technologies of Hadoop are HDFS for large data sets storage and Map Reduce programming model for large data sets computation.

From the perspectives of big data, synchronised harmonic analysis includes harmonic data storage, HSE and advanced analysis, which can be fulfilled by a Hadoop-based platform. The overall framework of the proposed platform and algorithm is depicted in Fig. 2. The upper-level block represents the data flow of the harmonics big data analysis process. Firstly, the raw data are collected from the harmonic voltage and current measurement points. Then, utilising the raw data, a HSE algorithm based on the MapReduce programming model is executed to achieve a full awareness of the distribution network. The results of HSE can be used for the further calculation of statistical indices on harmonics. Finally, the advanced analysis results can provide useful information about network harmonics situation to terminal users. The lower-level block presents the two core techniques used in the Hadoop platform. The raw data and intermediate data can be

Table 1: Harmonic data generated by individual instrument

| Data type | Harmonic order | Measurement period, s | Phase | Number of data per day |
|-----------|----------------|-----------------------|-------|------------------------|
| voltage amplitude | 1–50 | A/B/C | 4,320,000 |
| voltage phase | 1–50 | A/B/C | 4,320,000 |
| current amplitude | 1–50 | A/B/C | 4,320,000 |
| current phase | 1–50 | A/B/C | 4,320,000 |
divided into small data blocks and stored in the HDFS. The HSE and advanced statistical analysis algorithms are implemented using MapReduce programming model.

The harmonics analysis framework proposed in this paper has two kinds of potential application scenarios. The first one is off-line historical data processing. In this scenario, a huge amount of historical raw harmonics data could be processed rapidly and effectively by the proposed framework; then some further advanced analysis could be carried out based on the processed intermediate data. The second scenario is the on-line application, where high accuracy, the proposed framework can process the huge harmonics historical raw harmonics data could be processed rapidly and advanced analysis could be carried out based on the processed intermediate data. The second scenario is the on-line application, where high accuracy, the proposed framework can process the huge harmonics intermediate data. The second scenario is the on-line application, where high accuracy, the proposed framework can process the huge harmonics historical raw harmonics data could be processed rapidly and advanced analysis could be carried out based on the processed intermediate data. The second scenario is the on-line application, where high accuracy, the proposed framework can process the huge harmonics historical raw harmonics data could be processed rapidly and advanced analysis could be carried out based on the processed intermediate data. The second scenario is the on-line application, where high accuracy, the proposed framework can process the huge harmonics intermediate data. The second scenario is the on-line application, where high accuracy, the proposed framework can process the huge harmonics historical raw harmonics data could be processed rapidly and advanced analysis could be carried out based on the processed intermediate data. The second scenario is the on-line application, where high accuracy, the proposed framework can process the huge harmonics data timely to meet the requirements in practical applications.

In the next subsections, the algorithms of HSE and statistical analysis based on MapReduce programming model are presented and discussed.

3.2 Big matrix multiplication algorithm based on MapReduce programming for HSE

The MapReduce algorithm proposed in [46, 47] is used to deal with the issue of the multiplication of two big matrices in the study. For ease of explanations, the estimation matrix (2) can be rewritten as

\[
D = B \cdot C
\]

where \( D \) represents the estimation matrix \( X^* \) in HSE algorithm, while \( B \) and \( C \), respectively, represent the two big matrices \( A \) and \( Z_n \) in (2). The straightforward way of calculating matrix (3) is expressed by

\[
d(i, j) = \sum_{k=1}^{K} b(i, k) \cdot c(k, j)
\]

where \( d(i, j), b(i, k) \) and \( c(k, j) \) represent the elements in each matrix.

However, since \( B \) and \( C \) are big matrices, there are too many elements to be calculated by (4), which may result in the overload of a single computing node. Therefore, one possible solution is to split \( B \) and \( C \) into small blocks (sub-matrices) so that the multiplication of pairs of blocks can be handled by many single nodes parallelly.

The block of matrix \( B \) is defined as \( B[i_b, j_b] \), the block of \( C \) is defined as \( C[k_b, j_b] \) and the block of \( D \) is then defined as \( D[i_b, j_b] \), which can be calculated by

\[
D[i_b, j_b] = \sum_{k_b=1}^{k_b} N_{KB} \cdot D[i_b, k_b, j_b]
\]

where \( d(i, j), b(i, k) \) and \( c(k, j) \) represent the elements in each matrix.

\[
= \sum_{k_b=1}^{k_b} N_{KB} \cdot B[i_b, k_b] \cdot C[k_b, j_b]
\]

Based on the MapReduce framework, the reducers are responsible for the block multiplications. The mappers are in charge of distributing the block data to the reducers, with the assistance of a cautiously selected \( \langle \text{key}, \text{value} \rangle \) setting strategy.

In the proposed strategy, a single reducer, denoted as \( R[i_b, j_b] \), is used to calculate the block of \( D[i_b, j_b] \) according to (5). The reducer receives all the needed blocks \( B[i_b, k_b] \) and \( C[k_b, j_b] \) for \( 1 \leq k_b \leq N_{KB} \) from mappers. Then the reducer multiplies all the \( B[i_b, k_b] \) and \( C[k_b, j_b] \) block pairs and sums up the results, i.e.

\[
\sum_{k_b=1}^{k_b} N_{KB} \cdot B[i_b, k_b] \cdot C[k_b, j_b]
\]

A maximum of \( N_{IB} \cdot N_{JB} \) reducers multiplying blocks in parallel can be obtained.

The mappers are deployed to route a copy of each \( B[i_b, k_b] \) block to all of the \( R[i_b, j_b] \) reducers by indexing \( j_b \) such that \( 1 \leq j_b \leq N_{IB} \). Under the worst case, there are \( N_{IB} \) copies of matrix \( B \) and a total of \( N_{IB} \cdot K \cdot \langle \text{key}, \text{value} \rangle \) pairs. Likewise, the mappers are deployed to route a copy of each \( C[k_b, j_b] \) block to all the \( R[i_b, j_b] \) reducers by indexing \( i_b \) such that \( 1 \leq i_b \leq N_{IB} \). There are \( N_{IB} \) copies of matrix \( C \) for a worst case and total of \( N_{IB} \cdot K \cdot \langle \text{key}, \text{value} \rangle \) pairs. Thus with the strategy, a worst-case grand total of \( K \cdot (N_{IB} + 1 + N_{IB} \cdot J) \langle \text{key}, \text{value} \rangle \) pairs shall be transferred to the sort and shuffle phase of the MapReduce job. With the MapReduce structured matrix multiplication algorithm, the original bottleneck of calculating HSE over a huge dataset can be overcome by splitting the job into many map and reduce tasks which can be easily completed in parallel across clusters of commodity hardware.

3.3 Advanced harmonic statistics based on MapReduce programming

Further, harmonic analysis requires statistical calculations of enormous amounts of intermediate data based on HSE analysis. This problem can be also considered as a big data problem that can be solved by means of a similar way based on MapReduce programming model.

As an example, the calculation of the index of total harmonics distortion in (6) can be transformed into a MapReduce-based programming scheme

\[
\text{THD} = \sqrt{\frac{\sum_{i=1}^{N} (d(i)/I_i)^2 \times 100\%}{1}}
\]

\[\text{Step 1: Normalisation of the harmonics data. To calculate the THD index, the harmonics data obtained from HSE shall be normalised at first.}

\[\text{For each time } t, \text{ divide the } k\text{-th-order harmonic current } I_{k,t} \text{ by the magnitude of the fundamental current } I_{1,t}. \text{ Then all the harmonics data is normalised by } I_{1,t}/I_{1,t}.

\[\text{Step 2: MapReduce scheme for THD calculation. The average THD for bus } n \text{ is defined as}

\[
\text{THD}_{\text{Avg},n} = \sqrt{\frac{\sum_{i=1}^{N} \sum_{k=1}^{N} (I_{k,t}^n/I_{1,t}^n)^2}{N}} \times 100\%
\]

where \( T \) is the total number of time segments.

The task map aims at calculating the sum of relative harmonic current squares, i.e. \( \sum_{i=1}^{N} \sum_{k=1}^{N} (I_{k,t}^n/I_{1,t}^n)^2 \) with the key = \( n \), value = \( I_{k,t}^n/I_{1,t}^n \) in the \( \langle \text{key}, \text{value} \rangle \) pairs. The reduce task is responsible for calculating \( \text{THD}_{\text{Avg},n} \).

4 Case study and discussions

In this section, IEEE 14-bus system is studied to demonstrate the effectiveness of the proposed Hadoop framework. The performance of the HSE based on MapReduce programming model is compared with the traditional sequential calculation from the aspects of both computation time and accuracy. The HSE results are also compared via a stochastic analysis. The MapReduce algorithm is extended to a Spark-based platform to verify the extensibility of the proposed method.

4.1 Experiment setup

The experiments are carried out in a workstation comprising of 32-core Intel Xeon CPU running at 2.6 GHz with 64 GB of physical memory. The Hadoop clusters environment is established by virtual machines (VMs) with Ubuntu operating systems using the software of Oracle Virtual Box. There are seven VMs each with four CPU cores in order to emulate a distributed computing environment with seven clusters. Therefore, the Hadoop-based platform consists of
one Namenode (also JobTracker) and, to the maximum, six DataNodes (also TaskTracker). The specific details of the experiment environment implementation are displayed in Table 2.

The Hadoop-based harmonics big data analysis framework is depicted in Fig. 2. The Hadoop algorithm is deployed on the platform with up to six VMs aiming at the best performance. The result is compared with a sequential platform with only one VM. Under the experiment setting, a maximum of 24 CPU cores can be used at the same time when 6 VMS are put into service which means that 24 map or reduce tasks can be executed in parallel at the same time. The total number of the map tasks is determined by the size of input data and the block size of the HDFS split data blocks. For example, if the input data is 1024 MB for one file and the block size is set to 64 MB, the input data is split into 1024/64 = 16 blocks, then the number of map tasks will be 16. The number of total reduce tasks can be configured by the Hadoop configuration file. In most of the experiments, the total number of reduce tasks is set as 24.

4.2 Raw data generation

Due to the lack of real synchronised harmonics data in distribution network, the raw data used in this research is generated by harmonic simulation method [48, 49].

The harmonics simulation is based on IEEE 14-node system with four harmonic sources added at buses 3, 6, 8 and 9 as shown in Fig. 3. Three-phase model is used in this case as well as in following extended case, but only balanced situation is considered for simplicity.

Harmonics data simulation requires modelling for two kinds of elements in the network. The first kind is harmonic sources which are modelled as a current source [48] whose kth-order injection current \( I_k \) is determined by the fundamental voltage \( U_1 \) and harmonic voltage \( U_k \), expressed as

\[
I_k = f_k(U_1, U_k) \quad (8)
\]

The other components in the network are modelled for harmonics simulation according to the study in [49]. Typical overhead lines can be modelled by a multiphase coupled equivalent-pi circuit using the positive sequence impedance data which is further simplified into a single-phase pi-circuit. Traditional generators are modelled as a harmonic impedance in harmonics network. Three-phase model in [49] is used to model transformers so that the phase-shifting effects can be considered in the network. Passive loads are modelled with parallel resistance and inductance.

With all the models for passive components in the network, the nodal admittance matrix \( Y_k \) for kth-order harmonic can be obtained. The relationship between harmonic voltage and current can be expressed as

\[
I_k = Y_k \cdot U_k \quad (9)
\]

Then the whole harmonic simulation process can be described as an iteration process as follows:

Step 1: Initialise fundamental voltage \( U_1 \) and harmonic voltage \( U_k \).

Step 2: Calculate the harmonic current according to (8).

Step 3: Calculate the active and reactive power \( P, Q \) of each harmonics sources. Then calculate fundamental voltage regarding harmonics sources as PQ nodes.

Step 4: Calculate harmonic current based on (8) and then calculate harmonic voltage based on (9).

Step 5: If the accuracy of harmonic voltage meets the requirement, then go to step 6; otherwise go back to step 4.

Step 6: If the accuracy of fundamental voltage meets the requirement, then end the iteration; otherwise go back to step 2.

Some random factors are considered in the outputs of generators and the energy consumption of the load in the distribution network, and some random noises following a normal distribution are added to the data.
distribution with a mean of 0 and a standard deviation of 10% calculated current or voltage value are added into the calculated value to simulate the errors of the harmonic measurement devices. And it should be noted that, for simplicity, the measurement deployment problem is not considered during the HSE computation. It is assumed all the harmonics data calculated during the simulation are used in the HSE as direct measurements or pseudo-measurements. Under this condition, the observability can be fully guaranteed and the data size can be big enough to test the ability of the proposed platform.

4.3 Evaluation of HSE and statistical results

To evaluate the efficiency and accuracy of the MHSE algorithm, a number of experiments are carried out based on the Hadoop platform by comparing with the sequential HSE (SHSE). There are 7.74 GB raw harmonics data in total and small blocks of the data with different size are intercepted from the raw data to evaluate the computation time for HSE with different sizes of input data. The comparison results are displayed in Fig. 5. The MHSE is conducted with three different sizes of input data by different number of VMs from 1 to 6. The MHSE shows a good scalability especially for larger input data. If 3.1 GB data is processed, the speedup by six VMs is 3.13 times than using one VM. However, when 7.74 GB data is processed, the speedup significantly increases to 3.97 times. The comparison results depicted in Fig. 5 indicate that the platform is quite suitable for big data analysis with the rapid increasing of dataset size.

The total number of map and reduce tasks will influence the performance of the MHSE. If the calculation is split into more tasks, the computation burden for each task is lighter. Due to the limitation of the number of CPU cores, only a maximum of 24 tasks can be run at the same time. Therefore, too many tasks may result in a negative influence on the total computation time. As mentioned above, the total number of map tasks is related to the block size of HDFS and the total number of reduce tasks can be selected equal to the number of CPU cores. The number of reduce tasks is closely equal to the CPU cores, all the reduce tasks can be launched at the same time after the map tasks are finished. In the study case, each CPU core is of the same computation capability and each reduce task is of almost the same computation burden, so the total number of reduce tasks is selected equal to CPU cores. The block size of HDFS varies from 32 to 128 MB so that the relationship between the total computation time and the number of map tasks can be studied accordingly. The results are listed in Table 3, demonstrating that the best choice is 64 MB block size with 40 map tasks for the studied case.

After the processing of MHSE, the harmonic distribution in the network is obtained. Some advanced analysis, for instance the calculation of THD, can be carried out base on the MHSE data. The THD of the harmonic analysis results obtained by the proposed Hadoop platform is compared with that of the raw data and shown in Fig. 6. It can be found that the difference between two results is relatively small. The average deviation of THD calculated based on MHSE data, which is defined as

$$\text{Deviation}_{\text{Average}} = \frac{1}{N} \times \sum_{n=1}^{N} \left| \text{THD}^n_{\text{MHSE}} - \text{THD}^n_{\text{Raw}} \right| \times 100\%$$

where $N$ is the total number of buses. The average deviation is 0.44% in this case, indicating significantly high accuracy. It should be acceptable considering the noise added into the raw data and the incomplete measure points.

Attributing to the characteristics of synchronisation and intensive measurements, synchronised harmonics have a great potential in providing comprehensive situation awareness of active distribution networks. However, as previously discussed, lack of the ability to store and process intensive data efficiently has become the dominating obstacle in exploiting synchronised harmonics analysis in distribution networks. The experimental case
in this chapter has proven that the proposed Hadoop-based framework is capable of conducting harmonic state estimation based on the synchronised harmonics data with relatively high efficiency and accuracy. Considering that the data size in practical implementation will be much larger than the experimental case, more benefits can be achieved from utilising the proposed Hadoop-based framework. Furthermore, abundant advanced analysis on the basis of harmonics, the THD as a preliminary example in this paper, can be integrated into this framework. These advanced harmonic analysis, for instance, harmonics situation visualisation, security check for renewable energy integration, power quality control and so on, can provide more comprehensive situation awareness and more customer-oriented services in the future active distribution networks.

4.4 Extended case study

To further validate the efficiency and scalability of the proposed framework, an extended case with large-scale distribution network is tested in this section. The experiment environment, the raw data generation process and the MHSE algorithm are all the same as the basic case. The test distribution network contains 1221 nodes, which is extended from the IEEE 123-node distribution system [50] by combining 10 IEEE 123-node systems at the feeder and adding several additional harmonics sources. The diagram of the extended test case is shown in Fig. 7.

As the large scale of the extended case, the harmonics data generated during an operational time interval (e.g. 15 min) is big enough to reveal the advantage of the proposed framework on processing huge data. Six data sets including 15, 30 min, 1, 5, 12 and 24 h harmonics data are selected to be processed by SHSE and MHSE separately. The volumes of the selected datasets range from 1.1 G (15 min) to about 80 G (24 h). The calculation time comparison is explicitly shown in Fig. 8.

It is shown from Fig. 8 that in a large-scale distribution network, even for 15 min data, the traditional SHSE method costs more than 3 times calculation time than the proposed MHSE method. When the time duration increases to 24 h, MHSE is about 6.5 times faster than SHSE. Similar with the results in the basic case, the computation time for MHSE increases much slower than SHSE with the increase of the input data size. The numerical results demonstrate the high computational efficiency of the proposed framework. It can have high potential for applications in practical active distribution networks, such as power quality situation evaluation for planning (based on historical harmonics data analysis), harmonics on-line monitoring, harmonics sources tracking and so on. The THD calculation is also implemented on the extended case. As the random noise added into the raw data during the data simulation process, there are slight differences between the THD results computed from the raw data and the intermediate data produced by the HSE algorithm. The deviation distribution for 24 h harmonics data is presented in Fig. 9, where the horizontal axis indicates the deviation interval and the vertical axis indicates the number of buses. The deviations of most buses are below 0.5% and the average deviation is 0.37%, demonstrating the high accuracy of the proposed framework.

5 Conclusion

A Hadoop-based platform for harmonics big data analysis is proposed in this paper. In addition, a harmonic state estimation algorithm and a THD statistical calculation algorithm both based on MapReduce programming model are implemented on the platform. Case studies based on IEEE 14-node system and
extended IEEE 123-node system are fulfilled. The experiment results show the significant efficiency the proposed MHSE algorithm by comparing with traditional SHSE. In addition, the computation times of the proposed MapReduce-based algorithm with respect to the number of VMs are investigated using different sizes of datasets. The accuracy of the proposed Hadoop-based platform is validated via comparing THD results on the basis of the MHSE data and reference data. In general, the paper proposes systematical platform to process and analyse the synchronised harmonics in distribution networks, with high potential of applications in smart grids.

6 Acknowledgments

This work was partially supported by National Natural Science Foundation of China (51761135015, 51261130472, 51577096), National High-Tech Research and Development Program (863 Program) of China (2014AA051901), International S&T Cooperation Program of China (2014DFG602670), EPSRC-NSFC Smart Grids Project (EP/K006475/1), China Postdoctoral Science Foundation (2015M580097), National Natural Science Foundation of China (5126130472, 51577096), and Hong Kong RGC Theme Based Research Scheme (T23-407/11-N). The work was supported by the Hundred Talents Program of Zhejiang University.

7 References

[1] Moghe, R., Lambert, F.C., Divan, D.: ‘Smart “stick-on” sensors for the smart grid’, IEEE Trans. Smart Grid, 2012, 3,(1), pp. 241–252
[2] Huang, J., Ge, S., Han, J., et al.: ‘A diagnostic method for distribution networks based on power supply safety standards’, Prot. Control Mod. Power Syst., 2016, 1,(1), pp. 1–8
[3] Nabavi, S., Zhang, J., Chakraborty, A.: ‘Distributed optimization algorithms for wide-area oscillation monitoring in power systems using interregional PMU-PDC architectures’, IEEE Trans. Smart Grid, 2015, 6,(5), pp. 2529–2538
[4] Albert, A., Rajagopal, R.: ‘Smart meter driven segmentation: what your consumption says about you’, IEEE Trans. Power Syst., 2013, 28,(4), pp. 4019–4030
[5] Peppanen, J., Reno, M.J., Thakkar, M.: ‘Parallel matrix multiplication algorithm with LAMP’, IEEE Trans. Power Electron., 2014, 19,(2), pp. 473–481
[6] Smart Electric Meters: ‘Advanced metering infrastructure, and meter communication: global market analysis and forecasts’, 2013. Available at https://www.navigantresearch.com/research/smart-meters
[7] Yu, N., Shah, S., Johnson, R., et al.: ‘Big data analytics in power distribution systems’. Proc. Innovative Smart Grid Technologies Conf. (ISGT), 2015, (1), pp. 1–5
[8] Xu, L.D., He, W., Li, S.: ‘Internet of things in industries: a survey’, IET Comput. Syst., 2014, (4), pp. 62–74
[9] Huang, A.Q., Crow, M.L., Heydt, G.T., et al.: ‘The future renewable electric energy delivery and management (FREEDM) system: the energy Internet’, Proc. IEEE, 2011, 99,(1), pp. 133–148
[10] Zhou, Z., Chawla, N.V., Jin, Y., et al.: ‘Grid intersection of renewable energy sources in the distribution level with power-quality improvement features’, IEEE Trans. Power Deliv., 2011, 26,(1), pp. 307–315
[11] Demirci, T., Kalaycioglu, A., Kucuk, D., et al.: ‘Wide-area oscillation monitoring systems’, IEEE Trans. Power Deliv., 2014, 29,(2), pp. 879–888
[12] Cao, Z., Lin, J., Wan, C., et al.: ‘Parallel matrix multiplication using MapReduce’. IEEE Ninth World Congress on Services Computing, 2014, (4), pp. 38–47
[13] Cao, Z., Lin, J., Wan, C., et al.: ‘Optimal cloud computing resource allocation for demand side management in smart grid’, IEEE Trans. Smart Grid, 2017, 8,(4), pp. 1943–1953
[14] Tsai, S.J.S., Luo, C.C.: ‘Synchronized power-quality measurement network with LAMP’, IEEE Trans. Power Deliv., 2009, 24,(1), pp. 484–485
[15] Carta, A., Locci, N., Muscas, C.: ‘GPS-based system for the Measurement of synchronized harmonic phases’, IEEE Trans. Instrum. Meas., 2009, 58,(3), pp. 586–593
[16] Laverty, D.M., Best, R.J., Brogan, P., et al.: ‘The OpenPMU platform for open-source power system measurements’, IEEE Trans. Instrum. Meas., 2013, 62,(4), pp. 701–709
[17] Chidurala, A., Saha, T.K., Mithulananthan, N.: ‘Harmonic impact of high penetration photovoltaic system on unbalanced distribution networks: learning from different photovoltaic network’, IET Renew. Power Gener., 2016, 10,(4), pp. 485–494
[18] Enslin, J.H.R., Keskes, P.M.: ‘A self-synchronizing instrument for harmonic source detection in power systems’, IEEE Trans. Instrum. Meas., 2005, 54,(11), pp. 13–23
[19] Chidurala, A., Saha, T.K., Mithulananthan, N.: ‘Harmonic impact of high penetration photovoltaic system on unbalanced distribution networks: learning from different photovoltaic network’, IET Renew. Power Gener., 2016, 10,(4), pp. 485–494
[20] Cui, Y., Lin, J., Wan, C., et al.: ‘Stochastic Bi-level trading model for an active distribution company with DGs and interruptible loads’, IET Renewable Power Gener., 2016, 11,(2), pp. 278–288
[21] Singh, M., Khadikar, V., Chandra, A., et al.: ‘Grid interconnection of renewable energy sources in the distribution level with power-quality improvement features’, IEEE Trans. Power Deliv., 2011, 26,(1), pp. 307–315
[22] Almeida, C.F.M., Kagan, N.: ‘Harmonic state estimation through optimal matrix multiplication’, 2009.
[23] Kersting, W.: ‘Radial distribution test feeders’. Proc. Power Engineering Society Winter Meeting, 2003, August 2003, pp. 908–912
[24] Medina, A., Cisneros-Magana, R.: ‘Time-domain harmonic state estimation’, IET Renew. Power Gener., 2011, 5,(3), pp. 1701–1709
[25] Muscas, C., Pilo, F., Pisano, G.: ‘MLlib: machine learning in apache spark’, IET Comput. Syst., 2014, (1), pp. 307–315
[26] Demirci, T., Kalaycioglu, A., Kucuk, D., et al.: ‘Wide-area oscillation monitoring systems’, IEEE Trans. Power Deliv., 2014, 29,(2), pp. 879–888
[27] Wache, M., Murray, R.: ‘Optimal meter placement for synchronized phasor measurements for distribution networks’, IEEE Power Energy Society General Meeting PESGM, 2011
[28] Borghetti, A., Nucci, C.A., Paolone, M., et al.: ‘Synchronized phasors monitoring during the inauguration of an active distribution network’, IEEE Trans. Smart Grid, 2011, 2,(1), pp. 82–91
[29] Norstad, J.: ‘A MapReduce algorithm for matrix multiplication’. 2009. Available at http://www.norstad.org/matrix/multiply/
[30] Zheng, J., Zhang, L., Zhu, R., et al.: ‘Parallel matrix multiplication algorithm based on vector line combination using MapReduce’, IEEE Nineteenth World Congress on Proc. Services (SERVICES), 2013, 28 June 2013–3 July 2013, pp. 193–200
[31] Chen, X., Lin, J., Wan, C., et al.: ‘Optimal meter placement for distribution network state estimation: a circuit representation based MILP approach’, IEEE Trans. Power Syst., 2016, 31,(6), pp. 4357–4370
[32] Norstad, J.: ‘A MapReduce algorithm for matrix multiplication’. 2009.
[33] Bonner, A.: ‘Modeling and simulation of the propagation of harmonics in electric power systems’, IEEE Power Systems Concepts, models, and simulation techniques IEEE Trans. Power Deliv., 1996, 11,(1), pp. 452–465
[34] Kersting, W.: ‘Radial distribution test feeders’. Proc. Power Engineering Society Winter Meeting, 2003, August 2003, pp. 908–912