21, rue d'Artois, F-75008 PARIS
http : //www.cigre.org

CIGRE US National Committee
2017 Grid of the Future Symposium

Cloud Based Analytical Framework for Synchrophasor Data Analysis

P. ETINGOV, Z. HOU, H. WANG, H. REN, D. ZARZHITSKY
PNNL
USA

J. DE CHALENDAR
Stanford University
USA

D. KOSTEREV, A. FARIS, S. YANG
BPA
USA

SUMMARY

The paper presents initial results from the development of a cloud-based, Big Data analysis platform for power systems. The computational pipeline uses the Apache Spark framework running in an OpenStack cloud infrastructure. A real-world phasor measurement unit (PMU) dataset has been used to carry out the analysis. Actual examples of power system events detection using synchrophasor data are presented. It has been shown that applications of the cloud based computing environment and the Apache Spark framework enable a significant increase in the computational efficiency of large-scale PMU data analysis.

KEYWORDS

PMU, machine learning, Apache Spark, Big Data, event detection.

pavel.etingov@pnnl.gov

INTRODUCTION

Rising deployments of phasor measurement units (PMUs), smart meters, digital fault recorders (DFRs), and other contemporary measurement devices dramatically increase the size of data collected by electrical utilities [1]-[3]. This digital information is frequently unstructured, has different time scales, and is stored on different servers and databases. The size of the collected datasets is growing rapidly, which complicates data processing and analysis. However, because the collected information contains many insights about the power system's state and its dynamic behavior, extracting this knowledge can significantly increase situational awareness, detect system-wide or local anomalies (e.g., under-frequency or voltage events), validate system models, and discover/predict equipment malfunctions.

For the past decade, technologies for Big Data analytics, cloud computing and machine learning (ML) have been developing very rapidly, and have been applied in many different engineering areas, including power system studies [4]-[6]. New methods and computer frameworks for Big Data collection and analysis are based on distributed storage and parallel processing of information. Many of the Big Data analytical frameworks are open-source software projects, making it possible to apply this technology to an organization's existing commodity computing infrastructure without incurring new licensing costs. The per-unit cost of hardware components (e.g., central processing units, memory, and storage) has decreased dramatically as a function of computational performance. Combined with the lack of licensing costs for the state-of-the-art analytic solutions for Big Data analysis, it allows for use of either on-premise or Infrastructure-as-a-Service (IaaS) computer clusters by a broader community of researchers and industrial customers. In addition, popular commercial cloud services offered by multiple providers (e.g., Amazon Web Services, Google Cloud, Microsoft Azure, etc.) take this abstraction even further with the Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) business models, which feature a high-degree of customization and a variety of pre-packaged solutions for both data management, analytics, visualization, long-term secure storage, and many other operational and mission-impacting concerns.

The Apache Hadoop [7] framework has been successfully used as a foundation by many software solutions for distributed data analysis. An active community of Hadoop developers produced a thriving open-source software ecosystem for high-performance data analysis. It distributes (i.e., partitions) large datasets across multiple storage and computation nodes within a computer network. Using many common Hadoop-inspired technologies, Apache Spark is a popular and widely-used open source framework for Big Data analysis and ML [8]. ML is a method of data analysis that automates analytical model-building. ML techniques could build general analytical models based on the data analysis and find hidden insights without being explicitly programmed for each specific problem. Moreover, an ML engine can continuously improve its model from new data. Spark is based on a high-performance distributed memory architecture and it achieves exceptional performance in parallel data processing. Together, Spark and Hadoop have been used in different areas including power system applications [9]-[10].

This paper presents initial results of synchrophasor information analysis conducted on a cloud-based Hadoop and Spark Big Data infrastructure. The framework is based on the Pacific Northwest National Laboratory's (PNNL) Institutional Cloud Computing OpenStack installation. The Hadoop Distributed File System (HDFS) is used to store the raw PMU information, and then Spark is used for data analysis and ML. Analysis results presented here are based on the real synchrophasor data provided by the Bonneville Power Administration (BPA). Several statistical and ML methods were developed and applied to this synchrophasor dataset to detect different types of events (e.g., frequency or voltage) and abnormalities. The aim of this work is to develop technologies and techniques that improve power system situational awareness and reliability.

SYNCHROPHASOR DATA ANALYTICAL FRAMEWORK

Computer cluster

PNNL's Institutional Cloud Computing system is based on the OpenStack open-source platform [11] for cloud computing, and the Hadoop and Spark computational environment is provided by the Cloudera Express distribution [12]. The computer cluster network topology used in this study is shown in Figure

1. It consists of 20 nodes including one master head node. Each node is equipped with eight core processors, 32 GB of RAM, and 100 GB of disk storage space. Apache Spark for Big Data analysis and ML [8], and the Apache Hive based structured query language (SQL) interface [13] for data storage, management are configured through the Cloudera Manager.

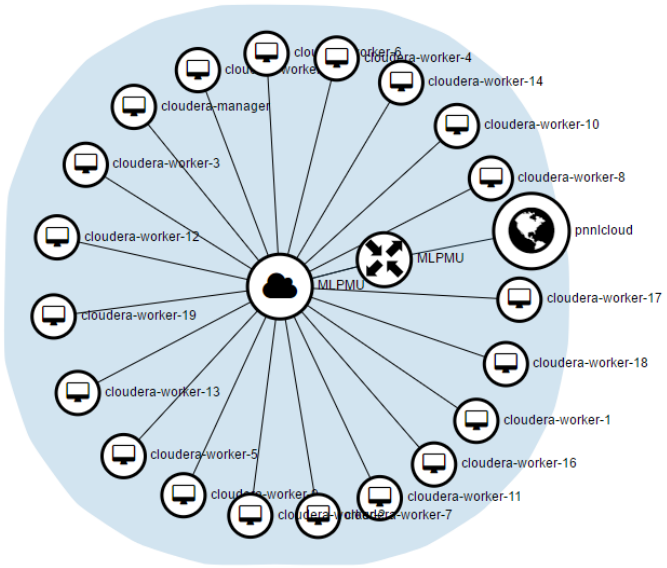

Figure 1. Computer cluster network configuration.

Our system's main functional components are diagrammed in Figure 2. PNNL receives the synchrophasor measurements as a real-time data stream from BPA, storing it at the PNNL's Electricity Infrastructure Operations Center (EIOC) [14] as a set of PDAT-formatted files. The PDAT format was developed by the BPA, and is used by the utility company to capture PMU measurements from multiple devices in binary files (the format is based on the IEEE Standard C37.118.2-2011 data frames) [15]-[16]. Each file contains one minute of PMU data, collected at the 60 samples per second rate. An approximate size of the dataset as a function of time is given in Table 1.

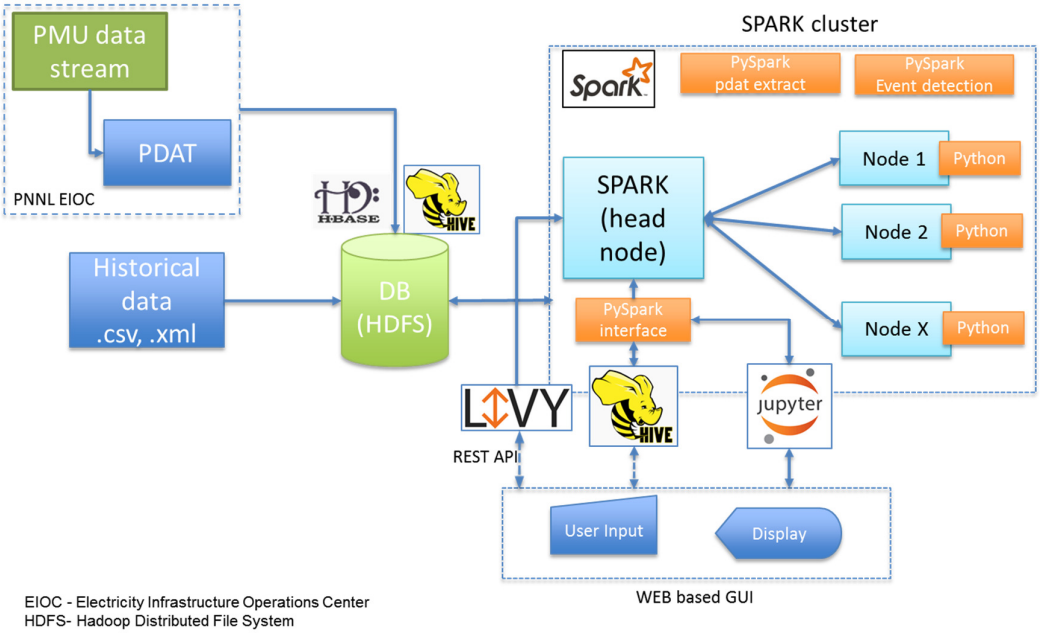

Figure 2. Cloud based framework.

Table 1. Approximate dataset size

1 minute	1hour	1 day	1 month	1 year
5 MB	300 MB	7.2 GB	216 GB	2.6 TB

Data extraction

All the PDAT files with the synchrophasor data are stored and distributed among the cluster nodes via the Hadoop Distributed File System (HDFS). The Python programming language, because of its wide use by the Data Science community and the availability of a large number of open-source data-processing modules, was selected as the programming environment for our data processing pipeline. The pipeline itself is split into several stages, and the interaction with the Spark execution engine is implemented using the PySpark binding. The first processing stage reads data from the HDFS hosted PDAT binary files and creates Spark data frames [8]. Here, the use of Spark enables significantly increased speed of data extraction (extraction of a one-hour dataset takes only 10-12 seconds compared to the 3-5 minutes required on a single personal computer).

As part of the second stage of data analysis, the Spark-processed data frames are saved as Hive tables in order to enable the use of Spark SQL application programming interface (API). Our design enables external modules, such as MS Windows standalone applications or web-based graphical user interfaces, to interact with Hive directly, further increasing the number of analytic and visualization options that can benefit from the cloud-based system architecture.

Events and anomaly detection

Several approaches for events and anomaly detection have been developed and implemented. In this paper, we present results of the event detection based on two approaches.

The first, relatively simple "engineering" approach is based on user-specified thresholds for signal values and duration (see Figure 3). This method is commonly used by electrical utilities to detect system events. To be qualified as an event, the signal should exceed a minimum or maximum threshold for a period of time longer than the minimum duration time. To avoid false alarms, cross validation signal checks (see Figure 4) for bad data detection, and dropouts are also used [17]. This approach is very efficient for system-wide events (e.g., an under-frequency event), which should be seen by all PMU devices located in different locations of the energy grid. For local events, a cross validation check (e.g., voltage drops) only obtain measurements from devices that are located in proximity to each other.

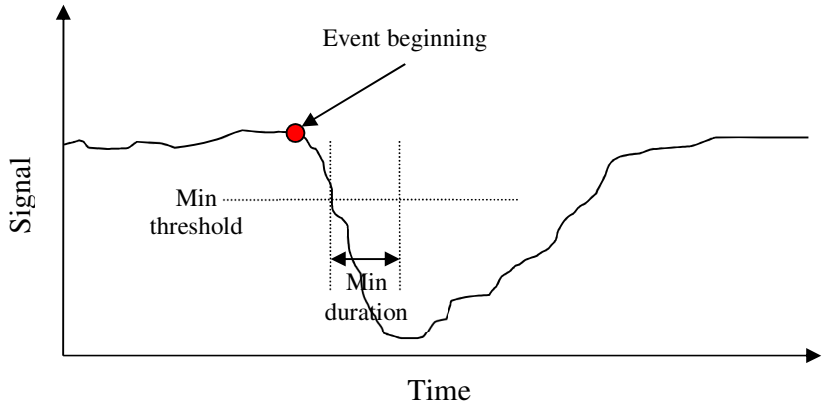

Figure 3. Event detection based on the threshold.

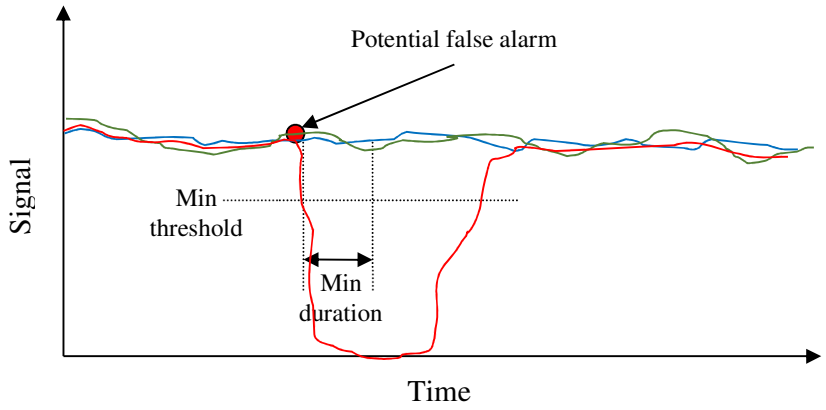

Figure 4. Cross signal check (system wide events) for bad data detection.

The second approach that has been developed and tested is based on the multiple resolution analysis (e.g., wavelet transform). Conventional anomaly detection can be accomplished by several common signal processing approaches, such as Eigenvalue Analysis (EA), Fourier Transform (FT), Short Time Fourier Transform (STFT), and Spectral Analysis (SA). With the large amount of PMU data, however, computational requirements can be demanding. Moreover, PMU signals are usually nonstationary and aperiodic, which makes the FT methods unsuitable for PMUs. Furthermore, STFT and SA approaches are based on fixed-sized windowing techniques, which are less accurate and less efficient in tracking signals in both time and frequency domains.

We adopted the wavelet transform approach, which separates one-dimensional signals into two-dimensional components that overlap in the time-frequency domain. Wavelet techniques have been widely used because of their ability to achieve multiple time-frequency resolutions [18]. Wavelet transforms proved to be very efficient [19] in signal analysis with the reduction of coefficient numbers as the scaling factor increases. Discrete wavelet transform (DWT) provides information in a computationally efficient manner, and is sufficient to decompose and reconstruct most power-quality problems. Wavelet-based anomaly detection has been successfully applied for detecting anomalies [20-21] for various real-world systems and problems. Wavelet-based multi-resolution analysis (MRA) uses wavelet function and scaling function to decompose and construct signals at various resolution levels, such that the anomaly phenomena can be detected and localized at each resolution level.

The proposed wavelet-based anomaly detection approach is shown in Figure 5. To localize anomalies at each decomposed resolution level, a moving-window-based outlier detection approach was developed. The moving window can retain both reliability and sensitivity of the detection performance. The anomalous score was set to be 1 if an anomaly was detected at each resolution level. The anomaly score matrices were the summation of scores at multi-resolution levels across all PMU devices installed in the system. These anomalous scores can be used to rank the 'likelihood' of abnormalities. Given the observation that the recorded actual events last between 5 and 20 seconds each, the detected anomalies were further pre-screened by the event durations.

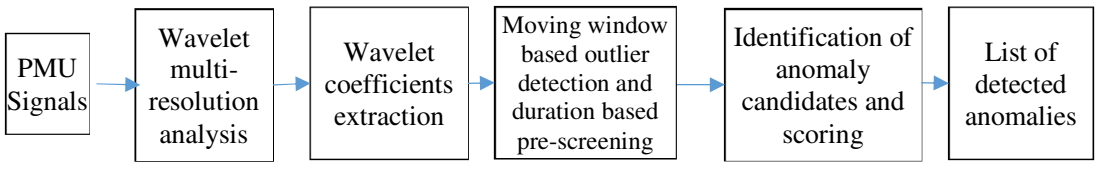

Figure 5. Wavelet based anomaly detection approach flow chart.

STUDY RESULTS

The developed detection framework was applied to the actual western interconnection synchrophasor data. Data collected for three months by 12 PMU devices was analyzed. The dataset includes voltage phasors and frequency time series. Application of the first approach allows us to detect multiple frequency events. An example of detected frequency events using "engineering" approach is presented in Figure 6. A minimum frequency threshold equal to 59.95 and minimum event duration 0.5 second was used.

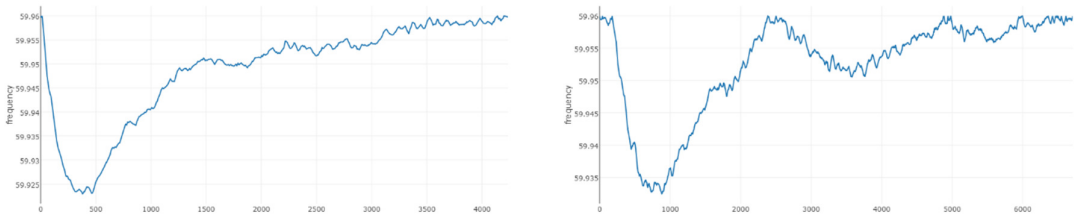

Figure 6. Example of the detected frequency events using "engeeniring approach".

Examples of frequency and voltage events detected using wavelet approach are given in Figures 7 and 8. One can see that requency events are system-wide and normally should be captured by all PMU devices. At the same time, voltage events have a different underlying nature. Figure 8a shows an event that is captured by all PMUs. At the same time, an event presented in Figure 8b can be clearly seen only from closely located PMUs.

Overall, the wavelet approach demonstrated better performance compared to the first method. The wavelet method detected all events captured by the "engineering" approach, and also detected some additional events not captured by the first approach. The list of detected events was compared with the "ground truth" event log provided by BPA, and all abnormal events from the log were correctly marked by the wavelet event detection approach. At the same time, the authors recognize that both methods presented in this paper require additional development and tuning in terms of event detection quality and computational performance. This improvement work is already in progress and will be reported in future publications.

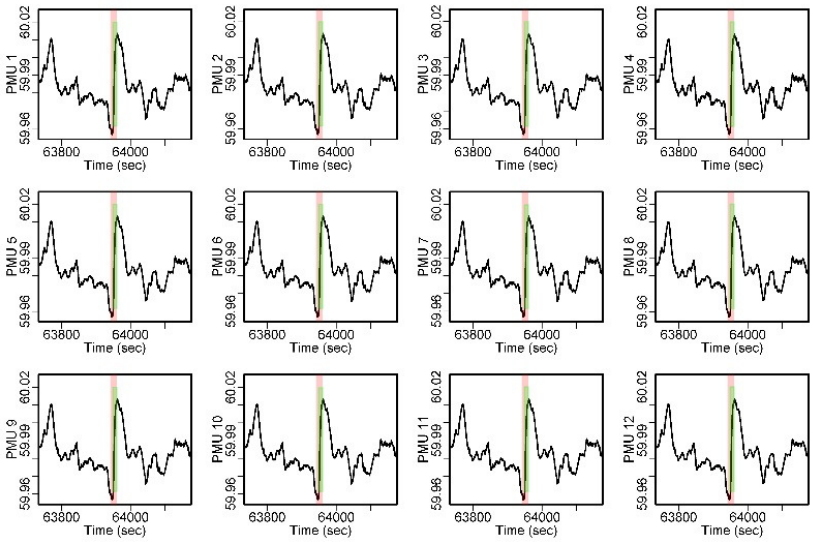

Figure 7. Example of detected frequency event using the wavelet approach.

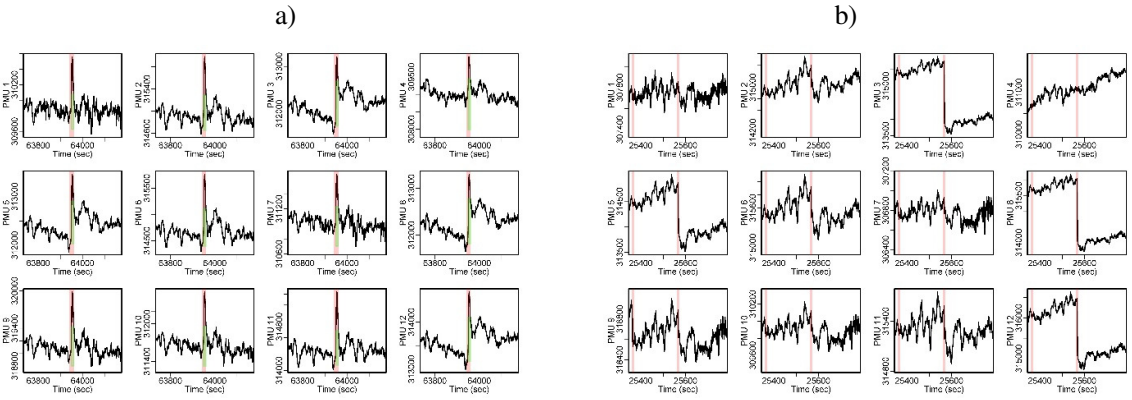

Figure 8. Example of detected voltage events using theh wavelet approach: a) system wide event; b) local event.

CONCLUSIONS AND FUTURE WORK

The framework for PMU data analysis based on the Apache Spark technology has been developed and tested using real system synchrophasor measurements. Software modules to efficiently read large volumes of PMU information from BPA PDAT files and pre-process raw data for event detection have been implemented using the Python programming language and the PySpark interface. A 3-month dataset was analyzed and multiple system events were detected. The proposed detection method based on a wavelet technique demonstrated superior performance compared to the commonly used engineering approach that relies on threshold exceedance triggers.

We showed that application of a cloud-based platform and Apache Spark significantly increases the computational throughput of the application. Analysis of several months of PMU data using the 20 node computer cluster with commodity hardware takes up-to several hours. By comparison, a similar job executed on a single personal computer can take several days to complete.

In future work, we plan to continue both a mathematical and software enhancement of this framework's functionality by adding new analytical modules and additional data sources, like supervisory control and data acquisition (SCADA) data, and also weather information. We are also going to increase the number of nodes and the size of memory in the research cluster, and improve performance of the analytical module by optimizing the software implementation of our analytic components.

ACKNOWLEDGEMENT

This work is supported by the U.S. Department of Energy (DOE) through the Grid Modernization Laboratory Consortium (GMLC) program.

BIBLIOGRAPHY

[1] M. Kezunovic, L. Xie and S. Grijalva, "The role of big data in improving power system operation and protection," 2013 IREP Symposium Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid, Rethymno, 2013, pp. 1-9. Institute of Electrical and Electronics Engineers, Piscataway, NJ.

[2] B. Wang, B. Fang, Y. Wang, H. Liu and Y. Liu, "Power System Transient Stability Assessment Based on Big Data and the Core Vector Machine," in IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2561-2570, Sept. 2016.

[3] D. Zhou et al., "Distributed Data Analytics Platform for Wide-Area Synchrophasor Measurement Systems," in IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2397-2405, Sept. 2016.

[4] J. Zheng and A. Dagnino, "An initial study of predictive machine learning analytics on large

volumes of historical data for power system applications," 2014 IEEE International Conference on Big Data, Washington, DC, 2014, pp. 952-959. Institute of Electrical and Electronics Engineers, Piscataway, NJ.
[5] T. Xia, "A Future Oriented Data Platform for the Electric Power Grids," CIGRE Grid of the Future Symposium, Chicago, Illinois, October 11-13, 2015. International Council on Large Electric Systems, Paris, France.
[6] V. A. H. Vajjala, "A novel solution to use Big Data technologies and improve demand response program in aggregated residential houses," 2016 IEEE Conference on Technologies for Sustainability (SusTech), Phoenix, AZ, 2016, pp. 251-256. Institute of Electrical and Electronics Engineers, Piscataway, NJ.
[7] Apache Hadoop. (June 1, 2017). Apache Hadoop 3.0 Documentation. [Online]. Available: http://hadoop.apache.org/ Wakefield, MA.
[8] Apache Spark. (June 1, 2017). Apache Spark 2.1 Documentation. [Online]. Available: http://spark.apache.org/ Wakefield, MA.
[9] G. Zhou et al., "The static security analysis in power system based on Spark Cloud Computing platform," 2015 IEEE Innovative Smart Grid Technologies - Asia (ISGT ASIA), Bangkok, 2015, pp. 1-6. Institute of Electrical and Electronics Engineers, Piscataway, NJ.
[10] F. Bach, H. K. Çakmak, H. Maass and U. Kuehnapfel, "Power Grid Time Series Data Analysis with Pig on a Hadoop Cluster Compared to Multi Core Systems," 2013 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, Belfast, 2013, pp. 208-212. Institute of Electrical and Electronics Engineers, Piscataway, NJ
[11] OpenStack. (June 1, 2017). OpenStack User Guide. [Online]. Available: https://www.openstack.org/
[12] Cloudera. (June 1, 2017). Cloudera User Guide. [Online]. Available: https://www.cloudera.com Palo Alto, CA
[13] Apache Hive. (June 1, 2017). Apache Hive 2.1.1 Documentation. [Online]. Available: https://hive.apache.org, Wakefield, MA
[14] Pacific Northwest National Laboratory, Electricity Infrastructure Operations Center. [Online]. Available: http://eioc.pnnl.gov/, Richland, WA.
[15] A. Faris, "BPA Synchrophasor Lab Tools," WECC JSIS meeting, September, 2016. [Online] Available: https://www.wecc.bz/Administrative/18%20BPA%20SP%20LabTools-2016_09.pdf, Bonneville Power Administration, Portland, OR.
[16] IEEE, "IEEE Standard for Synchrophasor Data Transfer for Power Systems," IEEE Std C37.118.2-2011 (Revision of IEEE Std C37.118-2005), pp. 1-53, 2011. Institute of Electrical and Electronics Engineers, Piscataway, NJ.
[17] M. Wu and L. Xie, "Online Detection of Low-Quality Synchrophasor Measurements: A Data-Driven Approach," in IEEE Transactions on Power Systems, vol. 32, no. 4, pp. 2817-2827, July 2017.
[18] S. Mallat, A wavelet tour of signal processing: Academic press, 1999.
[19] J. J. Benedetto and S. Li, "The theory of multiresolution analysis frames and applications to filter banks," Applied and Computational Harmonic Analysis, vol. 5, pp. 389-427, 1998.
[20] S. Bruno, M. De Benedictis, and M. La Scala, "" Taking the pulse" of power systems: Monitoring oscillations by wavelet analysis and Wide Area Measurement System," in Power Systems Conference and Exposition, 2006. PSCE'06. 2006 IEEE PES, 2006, pp. 436-443. Institute of Electrical and Electronics Engineers, Piscataway, NJ.
[21] W. Lu and A. A. Ghorbani, "Network anomaly detection based on wavelet analysis," EURASIP Journal on Advances in Signal Processing, vol. 2009, p. 4, 2009.