Cloudwave: Distributed Processing of "Big Data" from Electrophysiological Recordings for Epilepsy Clinical Research Using Hadoop

Catherine P. Jayapandian, BS[1], Chien-Hung Chen, BS[1], Alireza Bozorgi, MD[2], Samden D. Lhatoo, MD, FRCP[2], Guo-Qiang Zhang, PhD[1], Satya S. Sahoo, PhD[1]
[1]Division of Medical Informatics, Case Western Reserve University, Cleveland, OH;
[2]Department of Neurology, Case Western Reserve University, Cleveland, OH

Abstract

Epilepsy is the most common serious neurological disorder affecting 50-60 million persons worldwide. Multi-modal electrophysiological data, such as electroencephalography (EEG) and electrocardiography (EKG), are central to effective patient care and clinical research in epilepsy. Electrophysiological data is an example of clinical "big data" consisting of more than 100 multi-channel signals with recordings from each patient generating 5-10GB of data. Current approaches to store and analyze signal data using standalone tools, such as Nihon Kohden neurology software, are inadequate to meet the growing volume of data and the need for supporting multi-center collaborative studies with real time and interactive access. We introduce the Cloudwave platform in this paper that features a Web-based intuitive signal analysis interface integrated with a Hadoop-based data processing module implemented on clinical data stored in a "private cloud". Cloudwave has been developed as part of the National Institute of Neurological Disorders and Strokes (NINDS) funded multi-center Prevention and Risk Identification of SUDEP Mortality (PRISM) project. The Cloudwave visualization interface provides real-time rendering of multi-modal signals with "montages" for EEG feature characterization over 2TB of patient data generated at the Case University Hospital Epilepsy Monitoring Unit. Results from performance evaluation of the Cloudwave Hadoop data processing module demonstrate one order of magnitude improvement in performance over 77GB of patient data. (Cloudwave project: http://prism.case.edu/prism/index.php/Cloudwave)

1. Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent, unprovoked seizures affecting variety of mental and physical functions. Epilepsy affects about 50-60 million persons worldwide and is the most common serious neurological disease. In the United States, the Centers for Disease Control and Prevention estimates that epilepsy affects about 2.2 million Americans with a greater prevalence than Parkinson's disease or multiple sclerosis [1]. Epilepsy patients experience seizures due to abnormal signaling by clusters of nerve cells in the brain, which may briefly alter a person's consciousness, movements or actions. Electrical activity of the brain, including seizure events, is recorded as electrophysiological data, such as electroencephalogram (EEG). Electrophysiological data is essential for diagnosis, drug medication, and long-term care in epilepsy patients. Specifically, EEG signal data is considered as gold standard for pre-surgical evaluation of epilepsy patients and helps in characterizing the location and extent of the epileptogenic network that is used to guide the surgical procedure [2].

EEG data represents postsynaptic potentials from large group of neurons and electrodes (either scalp or intracranial) are used to record the voltage differences between the different brain regions. Patients are usually admitted to Epilepsy Monitoring Units (EMU) to record multiple channel signal data, including EEG, heart rate, blood oxygen levels, and electrocardiogram (EKG), over a period of five days. These comprehensive multi-channel patient recordings (in the order of 100s) generate very large datasets, for example a 24-hour recording for a patient represents 8000 screen images with about 5-10GB of data is generated for a single patient. An EMU usually admits about 100-150 patients in a year, which creates significant data management challenges similar to other "Big Data" application in terms of efficient storage, visualization, and analysis [3] [4]. The increasing need for multi-center clinical research studies exacerbates this challenge by introducing the need to share, interoperate, and integrate signal data in real time.

In addition, collaborative access to electrophysiological data requires reconciling heterogeneous data formats, cross-platform applications, and integrated environment that keep track of process used to generate results for reproducibility [5]. Traditional approaches to signal data management involving standalone tools (e.g. Nihon Koden neurology software tool [6]) are not suitable for collaborative research. Researchers face a number issues in using standalone tools for managing large scale, multi-modal electrophysiological data, such as:

1. No support for synchronous visualization and interaction with shared electrophysiological datasets across institutions by multiple researchers;
2. Limited re-usability of tools across different institutions and projects due to heterogeneous computing environments, such as operating systems, hardware configuration, and software libraries; and
3. Difficulty in maintaining software deployed at multiple sites over the life cycle of the tool/application.

In contrast, Web-based applications not only address the above challenges with easy accessibility through use of ubiquitous Web browsers (e.g. Microsoft Internet Explorer), but also have the ability to be transparently integrated with cloud computing resources to support extreme scalability, high fault tolerance, and high rate of service availability with low cost [7]. Scientific data management tools are rapidly adopting the cloud computing paradigm, which involves Web-based access to both storage and computing resources [8]. The cloud-computing infrastructure is also ideally suited for managing "big data" in clinical settings with *strict data security and accessibility features*. For example, Amazon Web Services (AWS) and Microsoft Azure platform provide a reliable, scalable, and inexpensive computing platform "in the cloud" that can support health care customers' applications in a manner consistent with HIPAA and HITECH [9, 10].

Many clinical research projects involving big data resources can also take advantage of high performance distributed computing algorithms, such as the popular Map Reduce approach [11], to efficiently process very large datasets stored in "private cloud" using open source Hadoop implementation [12]. In this paper, we introduce *Cloudwave*, a Web-based, ontology-driven electrophysiological data analysis and visualization platform that uses Hadoop for processing large scale multi-channel signal datasets. Cloudwave has been developed as part of the National Institute of Neurological Disorders and Strokes (NINDS) funded multi-center Prevention and Risk Identification of SUDEP Mortality (PRISM) project, which is part of the NINDS SUDEP Centers Without Walls initiative. Sudden and Unexpected Death in Epilepsy (SUDEP) is a poorly understood phenomenon, where the mechanisms of death are unknown and effective prevention strategies are yet to be defined [13].

The PRISM project aims to recruit about 1200 patients from four participating EMUs at the Case Western Reserve University (CWRU) University Hospitals-Case Medical Center (UH-CMC), Ronald Reagan University of California Los Angeles Medical Center (RRUMC-Los Angeles), the National Hospital for Neurology and Neurosurgery (NHNN, London, UK), and Northwestern Memorial Hospital (NMH Chicago). Hence, the primary informatics challenge in the PRISM project is to allow real time access to patient data from different institutions in a secure collaborative environment for clinical researchers. Cloudwave is part of this informatics infrastructure with specific focus on enabling researchers to seamlessly search, query, and visualize signal data annotated with clinical events for patient cohort identification.

2. Background

Scientific applications use cloud resources for both storage and computing in a variety of domains, including astronomy (e.g. Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) [8]), genome sequencing (e.g. BlastReduce [14]), and molecular dynamics simulation [15]. There has been limited use of cloud infrastructure in clinical research primarily due to concern about storing patient data on public clouds, which can be addressed by using de-identified datasets [7]. For example, the International Epilepsy Electrophysiology Portal (IEEG) stores de-identified electrophysiological data on the Amazon Simple Storage Service (S3) and uses a node on the Amazon Elastic Compute Cloud (EC2) as a server [5]. The IEEG project aims to allow multiple participants to upload de-identified patient data on the IEEG portal, which will store the datasets on S3 and allow users to download and further analyze the datasets using local tools. Dutta et al. described a similar effort to store electrophysiological data on Hadoop cluster [16]. In contrast to these efforts, Cloudwave is an integrated electrophysiological data management platform that not only stores signal data on Hadoop but also uses novel application programming interface (API) for distributed processing of signal data for real time interactive visualization.

Cloudwave draws on background knowledge of three resources. The first resource is the Epilepsy and Seizure Ontology (EpSO) that is used as the reference terminology in Cloudwave to ensure consistent interpretation of signal annotations and support query features in the user interface. The European Data Format (EDF) is the second resource that describes a common representation format of electrophysiological data and the third resource is the Apache Hadoop platform that implements the Map Reduce distributed computing algorithm.

Epilepsy and Seizure Ontology (EpSO). EpSO is an epilepsy domain ontology that models epilepsy syndromes, EEG signal patterns, both scalp and intracranial electrodes, their placement scheme, and detailed brain anatomy to correlate signal events with their location [17]. EpSO uses the World Wide Web Consortium (W3C) recommended Web Ontology Language (OWL2) to represent the terms, their properties, and define appropriate restrictions to

allow automated reasoning, such as subsumption reasoning over the ontology class structure. At present, EpSO has more than 1000 classes, including re-used classes from the Foundational Model of Anatomy (FMA) [18] and the Neural ElectroMagnetic Ontologies (NEMO) [19]. EpSO has been developed in close collaboration with epileptologists and members of the International League Against Epilepsy (ILAE) Classification and Terminology Commission (CTC) to ensure compliance with the ILAE epilepsy and seizure classification system recommendations [20].

European Data Format (EDF). EDF is the de-facto standard based on eXtensible Markup Language (XML) for recording EEG data in commercial equipment and facilitating data interoperability in multi-center research projects [21]. An EDF data file consists of a "header record" followed by "data record". The header record fields contain various metadata descriptions, including the date and time of the recording, the number of data records, and the number of signals. The data record section has the actual EEG data in form of consecutive fixed-duration segments (epoch) of multiple recordings (polygraph) consisting of multiple signals in a single chart. Cloudwave processes and stores EDF data files in a relational database for easier access and querying.

Apache Hadoop. Hadoop is an open-source implementation of the Map Reduce algorithm, which supports data-intensive distributed applications and was originally developed at Google Inc. for processing large-scale Web data [11]. The Map Reduce algorithm consists of two steps called *map* and *reduce*, which involving partitioning of compute intensive tasks into several disjoint tasks that can be executed in a distributed computing environment using commodity hardware. The Hadoop framework transparently provides both reliability and scalability to applications. It uses a high performance distributed file system called Hadoop Distributed File System (HDFS) to store and manage the large scale data [22]. Hadoop enables applications to access thousands of computing nodes and petabytes of data.

3. Methods

Cloudwave is a high-performance integrated signal analysis platform with an intuitive Web interface for use by clinicians and research staff members that is integrated with Hadoop-based computation module for distributed processing of large electrophysiological signal datasets (Figure 1 illustrates the high level system architecture of Cloudwave). The Cloudwave platform was implemented using agile software engineering approach with close and frequent interactions with users for rapid prototype development and feedback. Cloudwave uses Model View Controller (MVC) architecture with Ruby on Rails technology stack and an open source JavaScript charting library. In the following sections, we describe the development of different components of Cloudwave.

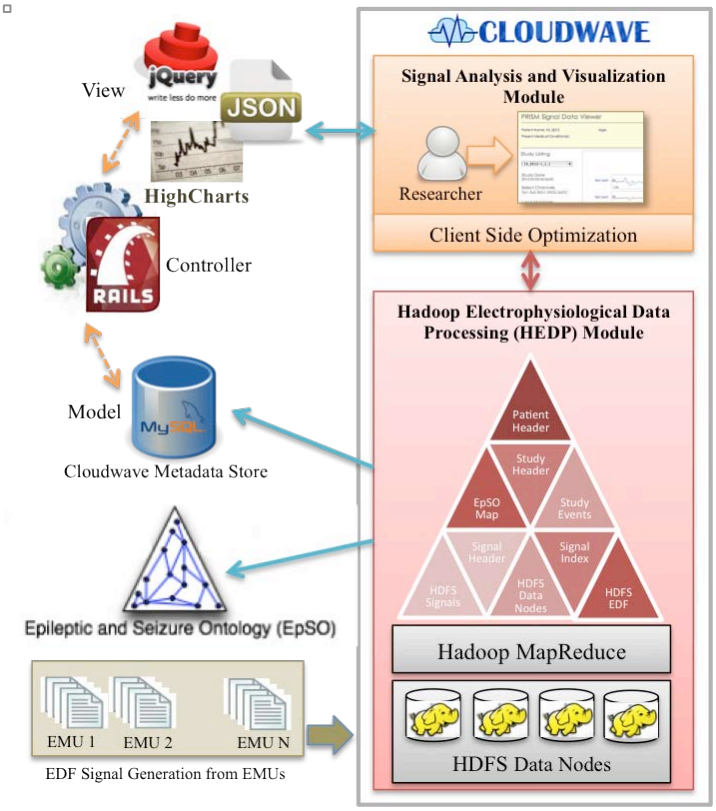

Figure 1: Cloudwave High Level Architecture

3.1 Hadoop Electrophysiological Data Processing (HEDP) Module

As we discussed earlier, electrophysiological data recorded for a patient during EMU admission spans a five-day period and generates multi-channel dataset. A typical recording involves 30-40 channels consisting of 20 EEG channels, 4 EKG channels, 1 channel for oxygen desaturation monitoring, 2 channels for respiratory signal, and 1 channel for monitoring blood pressure. Each participating EMU in the PRISM project generates the signal recordings as EDF files with each file corresponding to a single session of 6 hours of recording for a patient. Approximately, 20 EDF files are generated for each patient over a five-day admission period. As described earlier, an EDF file contains data

records with each data record containing 'duration' seconds of 'ns' signals, with each signal represented by a specific number of samples (described in the header). The 'duration' and 'ns' are specified in the header record, which is stored as Cloudwave metadata. For example, a six-hour recording with 30-40 signals is written in a EDF file according to a given time order: {*t1 – ch1 samples, ch2 samples, …, chn samples*} and {*t2 – ch1 samples, ch2 samples, …, chn samples*} represents two sets of signal data from channels *1* to *n* for time values *t1* and *t2*. The time duration is usually 0.1 sec for epilepsy patients with sampling rate ranging from 20-200 samples per 0.1 sec, therefore *t1=0.1sec and t2=0.2sec* in the above example. Due to this file structure, it is difficult to render signals and smaller signal time segments from an EDF file for analysis and visualization purposes. To address this challenge, individual signals are extracted from EDF files followed by signal analysis on a selected number of signal datasets (e.g. calculating RR intervals on ECG signal). The analyzed signals are stored in Cloudwave data format and directly accessed by the signal visualization module.

At present, the UH-CMC EMU has processed about 80 patients generating a total of 4000 EDF files with 400GB of data. This dataset is expected to grow rapidly as more patient datasets are aggregated from the other three participating EMUs in the PRISM project. Hence, it is essential to use distributed computing approaches to ensure scalability and acceptable performance during user interaction across multiple study centers. Unfortunately, there is no existing support for processing EDF files in Hadoop. To address this issue, Cloudwave defined and implemented a library of classes by extending the Hadoop API, namely:

1. ***EDFFileInputFormat****:* The *EDFFileInputFormat* class defines how the input EDF files are split and read by Hadoop. Given a directory of EDF files, the EDFFileInputFormat (extended from the abstract type FileInputFormat) read every EDF file as a key/value pair where the key is the filename and the value is the contents of the EDF file.
2. ***EDFRecordReader****:* This class facilitates the actual loading of data from its source (HDFS) using the input format and converts it into (*key*, *value*) pairs.
3. ***EDFMapper:*** The Mapper performs the first phase of the MapReduce program. A map task is assigned for each EDF file. The Mapper obtains metadata information about study and signal header from Cloudwave metadata store to process individual signal bytes and generates results as (filename, EDFWritable) pairs.
4. ***EDFWritable****:* The *EDFWritable* object has two data fields namely signal identifier/label and signal value as an array of bytes. The number of elements in this array denotes the number of data records of each signal given in the signal header. Signal values are stored in temporal order for generating the right signal waveform.
5. ***EDFReducer:*** Each reducer instance receives a key that corresponds to the filename and an iterator over all the corresponding *EDFWritable* values to be written as individual signal files.

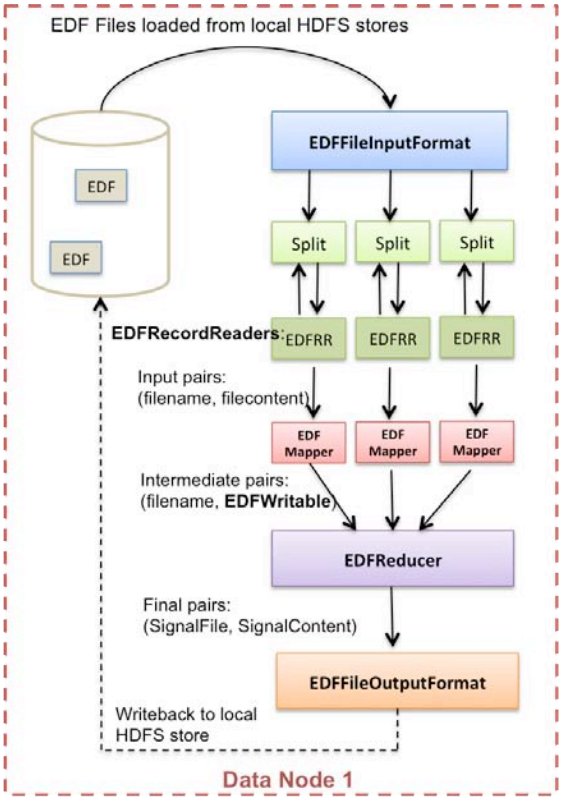

Figure 2: EDF Classes using Hadoop API

6. ***EDFFileOutputFormat****:* The class specifies the output format of the Reducer phase, to generate a files for every (key, value) with the key as the signal filename and the value as signal content.
7. ***EDFRecordWriter****:* The *EDFRecordWriter* creates a directory for each patient-study and writes the processed signal values into output files. For example, if an EDF file has 36 signals, then *EDFRecordWriter* will generate a directory with 36 files corresponding to 36 signals. Each file stores the data records of that specific signal.

Figure 2 illustrates the use of these new EDF-specific classes in the HEDP module to process the EDF. These classes will be released to the wider research community both as part of the PRISM project and to the Hadoop user

community members interested in electrophysiological signal processing. In the first phase of data processing in Cloudwave, the Header section of EDF files, which provides the structural, spatial and temporal information about the signal data, is extracted and stored in the Cloudwave metadata store. The Signal Visualization and Analysis (SVA) module (described in the next section) uses the signal metadata for query and retrieval of patient data. The metadata also contains information about the physical location of the actual signals in the HDFS, which is also used by the Data Visualization and Analysis module for loading and rendering the signal data. In the next phase, the EDF files are loaded into the HDFS for the Map and Reduce phases.

The HEDP module executes a Map Reduce "Job" (with *map* and *reduce* steps) for every patient corresponding to approximately 20 EDF files. In the *map* phase, the HEDP module loads the study files and assign each file to a *map* task. Each map task produces (*key*, *value*) pairs for the signals in the files as shown in Figure 3 for five patients used as examples in our current study. In the *reduce* phase, each *reduce* task consumes the fragment of (*key*, *value*) tuples assigned to it and writes the split files based on signals into HDFS. A directory is created for each patient with sub-directories for each channel. This is shown in the output of the *reduce* phase in Figure 3. The location of the processed files is updated in the Cloudwave metadata store for access by the SVA module.

The SVA module enable clinicians and research staff members to query, visualize, and analyze signal data in a Web-based interface using the output files from HEDP module.

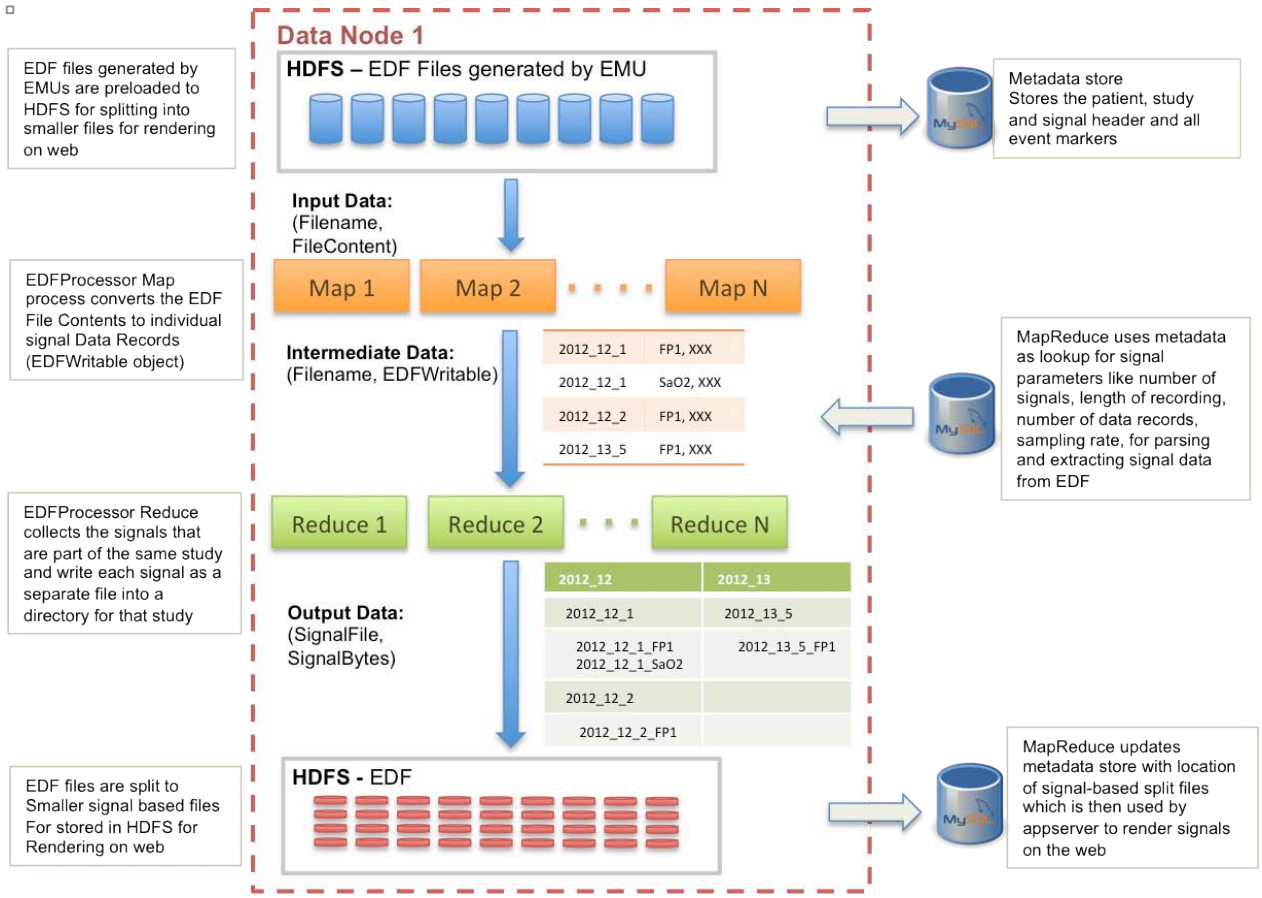

Figure 3: Cloudwave MapReduce Workflow

3.2. Signal Visualization and Analysis (SVA) Module

Clinicians create patient cohorts in the PRISM project using the characteristics of the electrophysiological signal data, for example:

1. Signal waveform corresponding to seizure events, such "Sign of Four" event, or "Onset of Jittery Phase";
2. The time duration of "EEG Suppression" in patients after a seizure occurrence;
3. Occurrence of cardiac arrhythmia through analysis of EKG signal; and

4. Time duration between “EEG Suppression” and “Return to Baseline”.

To support these queries in Cloudwave, the SVA module incorporates a rich set of query composition and visualization functionalities. The broad design goals of the SVA module were defined in close coordination with clinicians and staff members at the UH-CMC EMU and included a broad range of features, including:

a) Compressed rendering of long recordings in a signal page for ease of navigation;
b) Feature to browse select signal “segments” corresponding to seizure-related events without having to scroll through every page of recording;
c) Function to apply different “montage” (representation of EEG channels is referred to as a montage) settings on signals for feature extraction and analysis;
d) Ability to apply frequency filters to reduce noise and other signal artifacts; and
e) Allow increase or decrease of signal strength by applying amplitude filters to specific signals.

The SVA module is ideally suited for remote access as well as concurrent visualization and annotation of same signal data by multiple users. The SVA interface allows the user to select a patient study, render the signal with interesting events, and apply montage as well as filters. These options are implemented as simple drop down menus with ability to select multiple values in the drop down menu (Figure 4 illustrates the Montage, Channel and Study drop-down menus). These features are discussed in detail in the following sections:

3.2.1. Signal Montages. Once a patient study is selected, the user can select the appropriate montage, such as “bipolar montage” with waveform representing the voltage difference between two adjacent electrodes. Implementation of a montage requires computing the appropriate waveform from the relevant electrodes in real time in response to user query. SVA implements other montages including, “referential montage” (representing differences between a specific electrode and a designated reference electrode), “average reference montage” (with the outputs of all of the amplifiers aggregated and their average is used as the common reference for each channel), and “laplacian montage” (each channel represents the difference between an electrode and a weighted average of the surrounding electrodes).

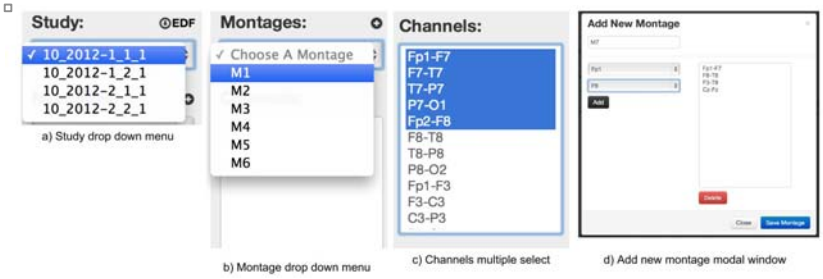

Figure 4: Study, Montage and Signal Selection

When a montage is selected, the channels defined by that montage are populated within a “select menu”. The user may select all or some of the signals/channels for display, as shown in part (c) of Figure 4. After selection of appropriate constraints, the SVA module will automatically display the relevant signals in the “charting area”. Users may also choose to create their own montages using the “add montage” interface, as shown in part (d) of Figure 4. The user has the ability to create any combination of signals for custom montages.

3.2.2. Event Selection. The “charting area” in the SVA module is implemented using an open source charting

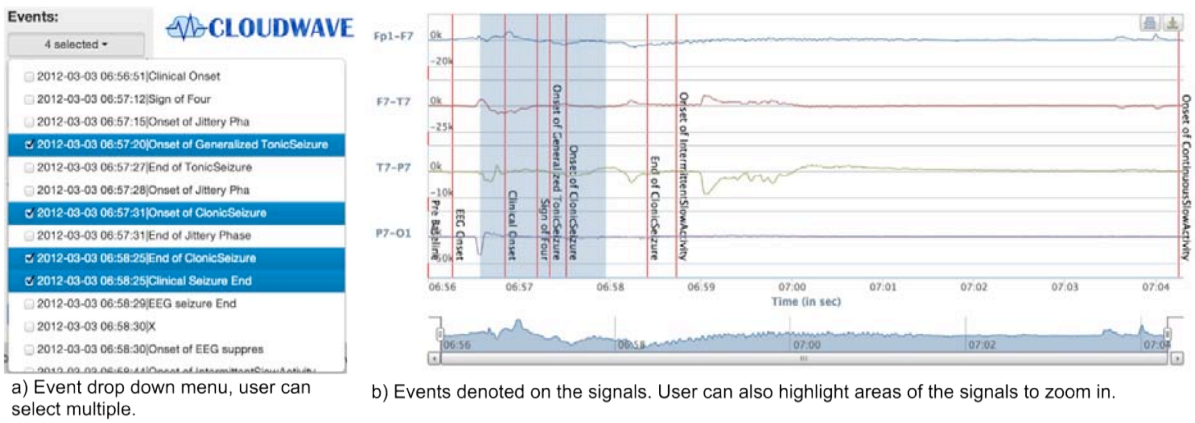

Figure 5: Selection of Events

library called Highcharts [23], which allows creation of general timeline charts in the JavaScript language, including sophisticated navigations options like a small navigator series, preset date ranges, date picker, scrolling and panning. The SVA module generates a multimodal chart for the selected signals that are synchronized temporally in a single screen (Figure 5). In addition, users can choose to view events that are marked on the signals to better navigate through the data. The ontology terms from EpSO are used to reconcile differences in the terminology used for describing seizure events across the different participating centers. Cloudwave uses REXML as a XML processor in Ruby for writing and reading the EpSO OWL file. The ontology-driven approach in Cloudwave enables rendering of the correct signal data segment with standardized event markings in response to user query. For example, EMU datasets can be labeled with either "Intermittent Seizure" or "Intermittent Slow Activity" to represent the same event, which is reconciled to a standardized "IntermittentSlowActivity" term modeled in EpSO. EpSO enables Cloudwave to map variations of a term used across different EMUs to a standard reference term to facilitate interoperability of signal datasets.

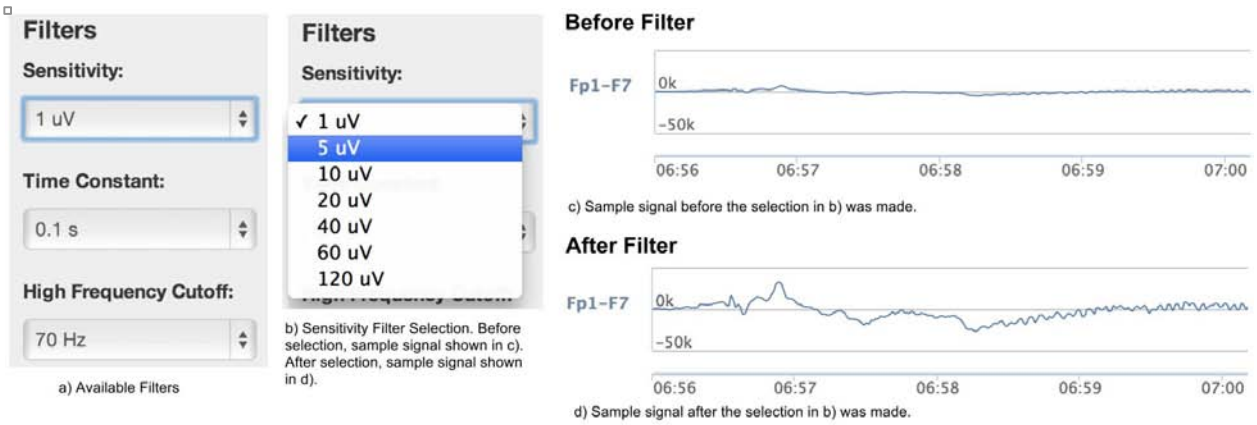

Figure 6: Selection of Filters

3.2.3. Filter Selection. Signal analysis is performed by the application of "sensitivity" and "frequency" filters as shown in part (a) of Figure 6. The sensitivity filter is calculated by multiplying the data by a value chosen from the drop down menu. The time constant filter is a low-cut filter that smooth out parts of the signals below a selected threshold frequency and the high frequency filter is a high-cut filter that smooth out arts of the signal above the given threshold frequency. Part (b) of the figure illustrates the selection of 5µV sensitivity filter, part (c) and (d) shows the signal before and after application of the filter.

3.2.4. Optimized Visualization. In addition to the Hadoop-based distributed file processing, the SVA module implements "client-side" optimizations to enhance the performance of signal visualization. For example, the signals are loaded asynchronously when possible and sent to the "client" interface in a format that allows "caching" to avoid repeated server access, which often slows down signal visualization. For signal rendering, best practices for JavaScript were followed to prevent memory leaks and eliminate unnecessary memory usage. The signals are rendered as they are selected, instead of waiting for the user to make all selection and submit the request, which results in notable reduction in the "wait time" associated with display of multi-graph chart.

In the next section, we describe the evaluation results of using Cloudwave to process electrophysiological data from patients admitted to the UH-CMC.

4. Results

Cloudwave was used to process electrophysiological data from 50 patients admitted to the UH-CMC EMU. The patient characteristics were tabulated that can be used by researchers in the PRISM project to query for patient cohorts (Table 1). Females outnumbered Males (61% vs. 39%) with a median age of 53 years within a range of 17-

Characteristics	Patients (% of 50)
Sex	
Male	39
Female	61
Age (Range: 17-75, Median: 53)	
0-20	8
21-40	52
41-60	34
61-80	6
Primary Diagnosis	
Seizure Semiology	
Epileptic Seizure	91
Non-Epileptic Seizure	9
Seizure Feature	
Lateralizing Sign	64
No Lateralizing Sign	
Epileptogenic Zone	
Focal	83
Generalized	17
Electrodes	
Scalp	98
Intracranial	2
Etiology	
Genetic Defect	9
Structural/Metabolic	13
Unknown	78
Medication	
Anti-epileptic	89
Anti-depressant	7
Neuroleptic	4

Table 1: Characteristics of patient enrolled in PRISM SUDEP study

75 years. Majority of the patients suffered from epileptic seizures (91%) and only a small number of patients had non-epileptic seizures (9%). Almost all the patient recordings were made using surface electrodes on the scalp (98%) and only 2% of the recordings were generated from intracranial electrodes. Intracranial electrodes generate significantly larger volume of signal data as compared to scalp electrodes (200 vs. 30-40) and the use of Hadoop platform enables Cloudwave to efficiently manage intracranial data.

4.1 Comparative Performance Evaluation. A comparative evaluation was performed to effectively measure the advantages of using Hadoop distributed computing platform for processing signal data in Cloudwave. The evaluation used five patient recordings collected from consented patients enrolled in the PRISM SUDEP study at UH-CMC EMU. The signal files were de-identified and manually verified to ensure removal of Protected Health Information (PHI) and stored in EDF data representation format. Table 2 shows the details of the dataset used in the evaluation with a total of 77GB of signal data. The evaluation involved processing of all five datasets on a standalone signal processing application running on a server machine with Quad-Core Intel Xeon 2.3 GHz processor, 3GB main memory, a 256KB L2 cache, and 8MB L3 cache. Cloudwave was installed on a single node cluster configuration with Intel Core i7 2.93 GHz processor, 16GB main memory, and 8MB cache.

DE identified Patient ID	Total Size in GB	Number of Studies	Number of Signals
2012_6	14.93	25	51-74
2012_7	15.01	26	51-73
2012_8	13.67	33	55-63
2012_13	18.1	50	42-74
2012_25	15.46	37	72

Table 2: Details of electrophysiological dataset used in comparative evaluation of Cloudwave

Two performance tests were performed. The first test compares the time taken (in minutes) to process data on the in standalone system and Cloudwave as the number of signals increase for 25 studies per patient. Figure 7 (a) shows that it takes 22-36 minutes for 10 signals and 1.5-3 hours (91-177 min) for 40 signals to be processed on the standalone system. In contrast, Figure 7 (c) shows that it takes only 4-6 min for 10 signals and 7-11 min for 40

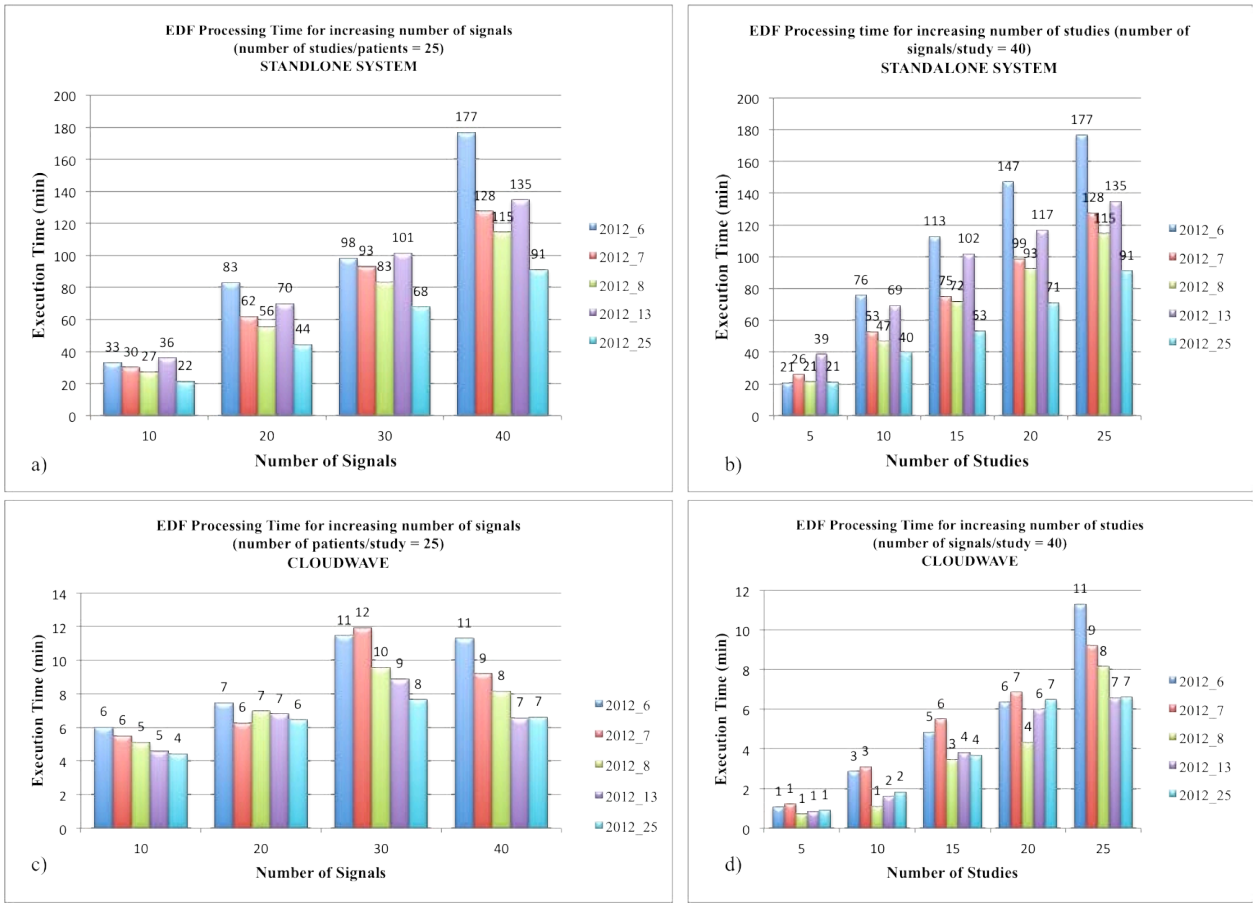

Figure 7: Comparative evaluation of performance of Cloudwave versus a standalone electrophysiological data processing application

signals on Cloudwave. All patient studies are six-hour recordings.

The second test was done to compare the execution time (in minutes) as the number of studies per patient is increased with each study consisting of 40 signals. Figures 7 (b) and (d) shows the results of this test. Figure 7 (b) shows that it takes 21-39 minutes for 5 studies and 1.5-3 hours (91-177 min) for 25 studies on the standalone system, but Cloudwave takes only 1 min for 5 studies and 7-11 min for 25 studies (Figure 7(d)).

Figure 8 (a) and (b) compares the average execution time on standalone system and Cloudwave for processing data from all 5 patients as the number of signals and studies are increased. The results clearly demonstrate that there is an order of magnitude difference between the average times taken by Cloudwave as compared to the standalone system for increasing number of studies having 40 signals/study.

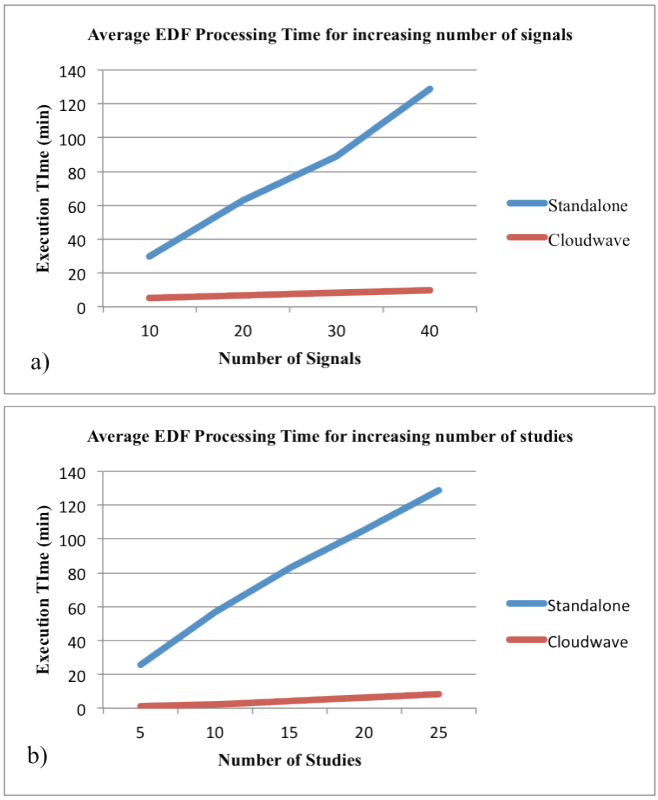

Figure 8: The average time taken by Cloudwave and the standalone application to process various categories of signal data

5. Discussion

Many Web-based applications that use "Big Data" resources often have low user satisfaction due to the slow response time in processing and rendering requested data on the user interface. Given the growing need for multi-center collaborative studies in clinical research using large datasets, such as electrophysiological datasets, there is an urgent need to adopt emerging computing platforms to meet these requirements. Hence, use of cloud computing resources that satisfy strict privacy and security requirements is a viable option for managing "Big Data" in clinical research. The results of the comparative performance evaluation of Cloudwave clearly demonstrate the significant advantages of using the Hadoop distributed computing platform for processing very large electrophysiological signal datasets.

This is important to address the need to implement interactive Web-based signal visualization and analysis functionalities for multi-center collaborative research in the PRISM project. In addition, Cloudwave is a generic, domain-agnostic signal management platform that can be used in a variety of medical disciplines that need to manage "Big Data", such as sleep medicine and neurodevelopmental disorders.

Limitations. As part of the next phase of development, we are implementing more complex signal processing algorithm, such as Heart Rate Variability (HRV), in Cloudwave to support greater number of analytical functionalities. In addition, the current version of Cloudwave does not support secure upload of signal data from remote locations. We are in the process of implementing a secured connection for uploading and sharing signal files from distributed locations through the SVA module.

6. Conclusions

Electrophysiological signal data, such as EEG, are often used as gold standard in the diagnosis and treatment of epilepsy. But, signal information generated during a patient's admission in an EMU results in very large size multi-modal datasets that cannot be managed using traditional standalone signal processing applications. This is especially important in case of multi-center collaborative clinical studies that require researchers to share and interact with signal data in real time. To address this challenge, we introduce the Cloudwave platform in this paper that features a Web-based intuitive signal analysis interface integrated with a Hadoop-based data processing module implemented on clinical data stored in a "private cloud". The Cloudwave SVA module provides real-time rendering of multi-modal signals with "montages" for EEG feature characterization of multi-modal patient data generated at the UH-

CMC EMU. Results from performance evaluation of the Cloudwave Hadoop data processing module demonstrate an order of magnitude improvement in data processing performance for more than 77GB of patient data. Cloudwave has applicability in a number of medical domains involving large-scale electrophysiological signal data processing and analysis.

Acknowledgement

This research was supported by the PRISM (Prevention and Risk Identification of SUDEP Mortality) Project (1-P20-NS076965-01) and in part by NIH Clinical and Translational Science Collaborative (CTSC) of Cleveland (UL1TR000439).

References

1. **Centers for Disease Control and Prevention** [http://www.cdc.gov/]. Accessed on July 24, 2013.
2. Rosenow F, Lüders H. **Presurgical evaluation of epilepsy**. *Brain* 2001, **124**:1683-1700.
3. Baraniuk RG. **More Is Less: Signal Processing and the Data Deluge**. *Science* 2011, **331**(6018):717-719.
4. Akil H, Martone ME, Van Essen DC. **Challenges and Opportunities in Mining Neuroscience Data**. *Science* 2011, **331**(6018):708-712.
5. **The International Epilepsy Electrophysiology Portal** [https://www.ieeg.org/]. Accessed on July 24, 2013.
6. **Nihon Koden Neurology** [http://www.nkusa.com/neurology_cardiology/]. Accessed on July 24, 2013.
7. Rosenthal A, Mork P, Li MH, *et al.* **Cloud Computing: A New Business Paradigm for Biomedical Information Sharing**. Technical Report. MITRE Corporation. 2010.
8. Nieto-Santisteban M, Simmhan Y, Barga R, *et al.* **Pan-STARRS: Learning to Ride the Data Tsunami**. In: *Microsoft eScience Workshop.* 2008.
9. **Amazon Web Services - Creating Healthcare Data Applications to Promote HIPAA and HITECH Compliance**. Technical Report. 2012. Accessed on July 24, 2013.
10. **Windows Azure HIPAA/HITECH Act Implementation Guidance**. Accessed on July 24, 2013
11. Dean J, Ghemawat S. **MapReduce: Simplified Data Processing on Large Clusters** In: *OSDI'04: Sixth Symposium on Operating System Design and Implementation: 2004; San Francisco*; 2004.
12. **Apache Hadoop** [http://hadoop.apache.org/]. Accessed on July 24, 2013.
13. Lhatoo SD. **Prevention and Risk Identification of SUDEP Mortality – the PRISM Project**. In. Case Western Reserve University: NIH-NINDS; 2011.
14. Schatz MC. **BlastReduce: high performance short read mapping with MapReduce**. Technical Report.
15. He C. **Molecular Dynamics Simulation Based on Hadoop MapReduce**. MS Thesis. University of Nebraska. 2011.
16. Dutta H, Kamil A, Pooleery M. *et al.* **Distributed Storage of Large-Scale Multidimensional Electroencephalogram Data Using Hadoop and HBase**. In: *Grid and Cloud Database Management* Edited by Fiore S, Aloisio, G.: Springer Berlin Heidelberg; 2011: 331-347.
17. Sahoo SS, Lhatoo, SD, Gupta DK, *et al.* **Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care.** *Journal of American Medical Association* 2013.
18. Barba C, Barbati G, Minotti L, *et al.* **Ictal clinical and scalp-EEG findings differentiating temporal lobe epilepsies from temporal "plus" epilepsies**. *Brain* 2007, **130**(7):1957-1967.
19. Dou D, Frishkoff G, Rong J, *et al.* **Development of NeuroElectroMagnetic Ontologies (NEMO): A framework for mining brain wave ontologies**. In: *Thirteenth International Conference on Knowledge Discovery and Data Mining (KDD2007): 2007; San Hose, CA*: ACM New York; 2007: 270-279.
20. Berg AT, Berkovic SF, Brodie MJ, *et al.* **Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005–2009.** *Epilepsia* 2010, **51**:676-685.
21. **European Data Format (EDF)** [www.edfplus.info/]. Accessed on July 24, 2013.
22. Shvachko K, Kuang H, Radia S, *et al.* **The Hadoop Distributed File System**. In: *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST).* NV; 2010: 1-10.
23. **Highcharts JS** [http://www.highcharts.com/]. Accessed on July 24, 2013.