A Match Made in Cloud? Meeting the Requirements of the Next Generation Neuroscience Research Using Configurable Cloud Infrastructure

Satya S. Sahoo[1,2], Cameron C. McIntyre[3], Samden D. Lhatoo[4]
[1]Division of Medical Informatics, [2]Electrical Engineering and Computer Science Department, [3]Department of Biomedical Engineering, [4]Department of Neurology, Case Western Reserve University, Cleveland, OH

Introduction

The increasing adoption of data-driven research in the healthcare domain is driven by two factors: (a) digitization of existing healthcare data to conform to federal laws (e.g. adoption of Electronic Health Records (EHRs)), and (b) use of sophisticated sensing and measurement instruments that record biomedical events at an unprecedented level of detail. In this paper, we use the neurosciences research domain to illustrate the computational challenges created by the increasing volume of multi-modal biomedical data generated at a high rate that needs to be interpreted in near real time (velocity), and is characterized by structural and semantic heterogeneity (variety).

The recently announced Brain Research through Advancing Innovative Neurotechnologies (BRAIN) initiative aims to undertake a comprehensive effort similar to the Human Genome Project to accelerate neuroscience research. Electrophysiological signal data together with imaging data are the two primary components of data-driven research in neurosciences and signal data is used as gold standard in patient care. For example, electroencephalogram (EEG) record brain activity and Electrocardiogram (ECG) measure cardiac events that are used for disease diagnosis and developing treatment strategy. The use of highly sensitive recording techniques using intracranial electrodes and precise placement approaches, such as Stereotactic EEG (SEEG), has led to rapid generation of large volume of signal data. About 20TB of data has been collected in our epilepsy monitoring unit (EMU) over the past three years. However, there are several computational challenges that need to be addressed before we can effectively leverage this biomedical Big Data, including:

1. **An extensible and configurable common storage layer** that supports new signal data partitioning approaches for scalable storage and subsequent analysis especially for complex nested signal data elements.
2. **Near real time access and analysis** of signal data for: (a) complex clinical event identification using scalable analysis algorithms (e.g. start or end of epilepsy seizures); and (b) efficient querying of signal annotations and related metadata for user queries and data mining applications.
3. **Interactive signal data visualization** by efficiently transferring large volumes of data over the network to Web browser interfaces for signal visualization, which can support collaborative multi-institution projects.

Current Approach: The Cloudwave Neuroinformatics Platform

We are developing a new cloud-based neuroinformatics platform called Cloudwave to address the data management challenges described in the previous section [1-4]. The Cloudwave project is developing the following resources:

1. **Real time data processing and analysis algorithms**: These algorithms use the MapReduce programming approach to scale with increasing volume and velocity of signal data.
2. **Efficient data query and retrieval using knowledge representation**: Using formal knowledge representation model called ontology for optimal data partitioning that support efficient query and data retrieval approaches. Ontology also allows reconciling data heterogeneity using common terminology and reasoning techniques.

3. **Flexible data representation format to support flexible storage schema in the cloud**: A common neuroscience data representation format that allows data to be stored in distributed file systems (e.g. Hadoop Distributed File System, HDFS) for parallel read and write tasks.
4. **Interactive signal data visualization**: A Web browser signal visualization interface that supports multiple signal filtering and query functionalities with interactive browsing capability.

Initial Results from the Cloudwave Project: Using the open source Hadoop implementation of MapReduce programming approach, the Cloudwave platform has developed scalable algorithms for clinical event detection. For example, the Cloudwave algorithms can compute cardiac events, such as instantaneous heart rate, an order of magnitude faster for single ECG channel data and 20 times faster for four ECG channel data as compared to existing approaches [3]. In addition, the Cloudwave data processing workflow generates partitioned segments of signal data that can be easily stored in HDFS and efficiently transferred to the visualization interface as compared to existing approaches [2, 4]. The Cloudwave uses the Epilepsy and Seizure Ontology (EpSO) to reconcile heterogeneity in signal metadata annotation for consistent interpretation and querying.

Programmable Cloud for Neuroinformatics: Vision and Roadmap

However, progress in the Cloudwave project has been impeded by the lack of a flexible cloud infrastructure that can be configured to support multiple requirements that are unique to neurosciences data, for example:

1. **Develop a tiered cloud-based storage infrastructure** for: (a) differential data access patterns and speed (e.g. specific signal data segments with clinical events need to be accessed in real time), and (b) supporting patient privacy requirements that comply with federal laws (e.g. HIPAA) and Institutional Review Board approvals;
2. **Support new parallelization and data partitioning approaches** to implement signal analysis algorithms that do not easily fit into the two-step MapReduce approach with recursive functions; and
3. **Enable use of domain ontology in cloud-based data processing and analysis** for efficient random access to specific segments of neuroscience data to support user queries and signal visualization applications.

The NSF Cloud program projects (Chameleon and CloudLab) can be used to develop a customized cloud infrastructure for neuroscience research that can address computational and storage challenges described above. The results of this experiment will be widely applicable to the biomedical and healthcare informatics community that need to leverage biomedical Big Data to advance research and patient care.

Reference:

[1] C. P. Jayapandian, Chen, C.H., Bozorgi, A., Lhatoo, S.D., Zhang, GQ, Sahoo, S.S., "Electrophysiological Signal Analysis and Visualization using Cloudwave for Epilepsy Clinical Research," in *The 14th MedInfo conference*, Copenhagen, Denmark, 2013, pp. 817-21.

[2] C. P. Jayapandian, Chen, C.H., Bozorgi, A., Lhatoo, S.D., Zhang, G.Q., Sahoo, S.S., "Cloudwave: Distributed Processing of "Big Data" from Electrophysiological Recordings for Epilepsy Clinical Research Using Hadoop.," in *American Medical Informatics Association (AMIA) Annual Symposium*, Washington DC, 2013, pp. 691-700.

[3] S. S. Sahoo, Jayapandian, C., Garg, G., Kaffashi, F., Chung, S., Bozorgi, A., Chen, C., Loparo, K., Lhatoo, S.D., Zhang, GQ, "Heartbeats in the Cloud: Distributed Analysis of Electrophysiological "Big Data" using Cloud Computing for Epilepsy Clinical Research," *Journal of American Medical Informatics Association (Special Issue on Big Data),* vol. 21, pp. 263-71, 2014.

[4] C. Jayapandian, Chen, C.H., Dabir, A., Zhang, G.Q., Lhatoo, S.D., Sahoo, S.S., "Domain Ontology As Conceptual Model for Big Data Management: Application in Biomedical Informatics," in *The 33rd International Conference on Conceptual Modeling (ER 2014)*, Atlanta, GA, 2014.