ICUWaveDB: A Big Data Approach to Capture and Processing Real-Time Streams in Critical Care

Lokesh K Chinthala, Akram Mohammed, Rishikesan Kamaleswaran
University of Tennessee Health Science Center

Abstract

Critically ill patients receive acute, complex and dynamic care within the intensive care unit (ICU), where they are monitored by a variety of biosensors that generate continuous data [1], [2]. Those sensors, such as the electrocardiogram (ECG), for instance, are sampled at hundreds of hertz (Hz). The capture and analysis of this data has been a challenge, due to the limited availability of methods and software to enable seamless capture, processing and analysis. While some open-source approaches exist for selected medical devices vendors [3], the availability of such software to perform preprocessing functions, including identifying noisy segments of the data and integrating clinical patient binding from hospital systems have been limited. We introduce an approach to support the preprocessing of real-time physiological data and stage such data in a NoSQL environment to enable scalable and efficient preprocessing.

Data Characterization

- Continuous waveform data is captured at a tertiary children's hospital from 11/ 2017, and is being stored in MongoDB [4] as a single collection
- The collection currently consists of 539,947,685 documents and occupies 3 terabytes (TB) of storage
- The structure of each document within the collection is hierarchal, and consists of metadata relating to the timestamp, room, device and sensor parameters (Figure 1). The data is stored in a delimited format
- This dataset as such do not contain any clinical identifiers and hence it is mapped to another dataset

(a) (b)

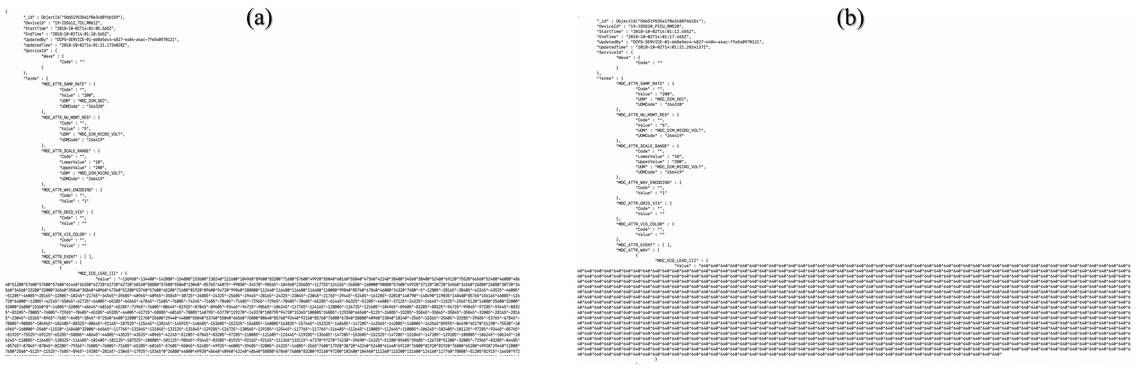


Figure 1: Sample waveform data without noise (a) and with noise (b)

Analytic Challenge

Continuous sampling of the waveform from the bedside introduces a number of artefacts, such as artificial amplifications, distortions, and null values. In the Draeger system, null values are represented by identical numeric, which may mimic real data. Here, we highlight a number of challenges in our approach to automatically identify null values in real-time data streams

Methodology

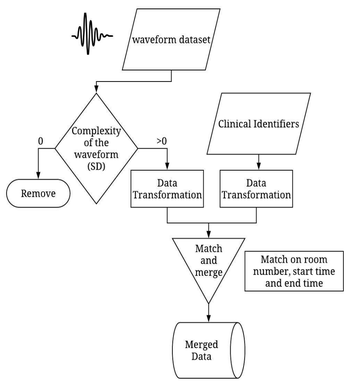


Figure 2: Flowchart of methods to support automated identification of noisy signals

Data Cleaning

- Null values in the data is flagged by calculating standard deviation of variable with waveform data in 'project' stage of MongoDB's aggregation framework

Data Transformation

- Before binding the waveform data with the clinical identifiers dataset, variables such as start time, end time and Device Id from both the datasets are transformed
- While the clinical identifiers dataset explicitly demarks patients by room and unit, the particular device identity is not recorded, conversely, the waveform dataset includes a device identifier but not the room number. Therefore, manual association is required from device to room

Time Synchronization

- While the waveform dataset is recorded using a standardized Coordinated Universal Time (UTC) timestamp, the clinical systems use local time. Therefore, synchronizing events are troublesome during seasonal events such as daylight savings time (DST), or errors through manual entry
- To overcome the first challenge, all the date variables were converted into the UTC format, with manual reviews performed during the DST transitions prior to the analysis

Patient Binding

- To assign clinical identifiers to the documents in waveform dataset, we implemented left outer join in 'lookup' stage of aggregation operation in MongoDB
- The datasets are matched and merged i.e., clinical identifiers are assigned to the documents in waveform dataset when the following three joining conditions are satisfied:
  - Room number of the clinical identifier dataset is equal to that of the waveform dataset
  - Start time of the waveform dataset greater than that of clinical identifier dataset
  - End time of the waveform data less than that of the clinical identifier dataset

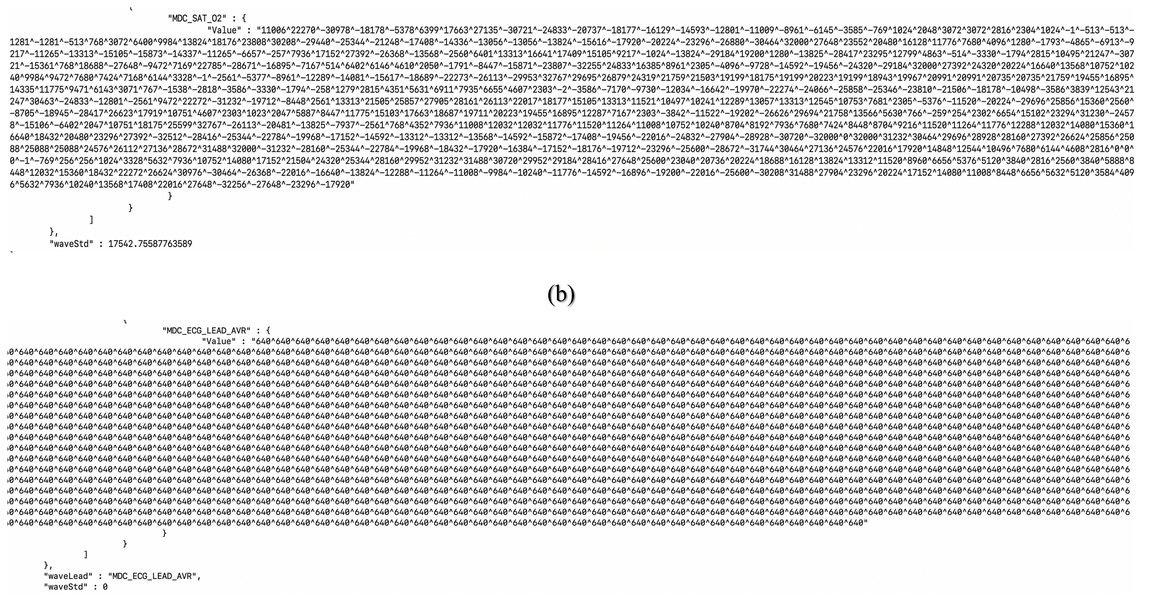
Results

Total size of the waveform dataset reduced from 3TB to 1TB after the data cleaning stage. With the reduction in collection size, the index size on the joining conditions also reduced and this improved patient binding performance exponentially

Document category	Documents (in million)
Total waveform dataset	**539.95**
Leads with Null values	**336.99**
MDC_ECG_LEADI	47.04
MDC_ECG_LEADII	32.02
MDC_ECG_LEAD III	42.15
MDC_ECG_LEAD_AVF	54.38
MDC_ECG_LEAD_AVR	54.38
MDC_ECG_LEAD_AVL	54.38
Wave-ECG_LEAD_AVR_INV	46.70
MDC_RESP	5.27
MDC_SAT_O2	0.67
% of Leads with Null values	**62%**

Table 1: Data reduction for each of the document category

(a)



(b)

Figure 3: Standard deviation for sample waveform data without noise (a) and with noise (b)

Conclusion

- Our methodology automatically identifies and cleans the real-time waveform data with Null values
- Using our methodology, we were able to efficiently reduce the data size to one-third of the original size

Future Work

The continuation of the current work is to create a publicly accessible Application Programming Interface (API) that can be used to clean the waveform data stored in MongoDB.

References

[1] G. C. Green, B. Bradley, A. Bravi, and A. J. E. Seely, "Continuous multiorgan variability analysis to track severity of organ failure in critically ill patients," Journal of Critical Care, vol. 28, no. 5, p. 879.e1–879.e11, 2013.

[2] R. Kamaleswaran, C. McGregor, and J. M. Eklund, "A method for clinical and physiological event stream processing," in Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, 2010, pp. 1170–1173.

[3] J. Plourde, D. Arney, and J. M. Goldman, "OpenICE: An open, interoperable platform for medical cyber-physical systems," in 2014 ACM/IEEE International Conference on Cyber-Physical Systems, ICCPS 2014, 2014.

[4] MongoDB Inc., "MongoDB: Open Source Document Database," 2018.

Acknowledgements

We would like to acknowledge Center for Biomedical Informatics at University of Tennessee Health Science Center, Baxter, Methodist LeBonheur Healthcare, Advanced Computing Facility at Oak Ridge National Laboratory.