### A Scalable, Secure and Realtime Healthcare Analytics Framework with Apache Spark

1. Md. Rezaul Karim, Ratnesh Sahay and Dietrich Rebholz-Schuhmann
2. The Insight Centre for Data Analytics, National University of Ireland Galway
3. {rezaul.karim, ratnesh.sahay, rebholz}@insight-centre.org

**Abstract**

*A Big Data analytics framework with related computing technologies can process huge amounts of real-time data to obtain tremendous insights for effective clinical decision making in the healthcare research. In this paper, we propose a healthcare analytics framework with Apache Spark for realtime healthcare monitoring and related analytics by integrating heterogeneous data sources like Internet of Things (IoT) enabled medical devices, Electronics Health Records (EHRs), picture* archiving and communication system (PACS)*, wearable sensors, and public open health datasets. The ultimate goal of this research is to provide an effective, scalable and secure framework for increasingly important public health surveillance.*

**1. Introduction**

Tremendous use of biomedical devices and data sources including wearable sensors, IoT enabled devices, EHRs, medical health records (MERs), PACS, public open health data (POD), Linked Opened Data (LOD) and medical knowledge from the biomedical research literatures are creating a new and promising area for applying Big Data healthcare analytics [5]. These make Apache Spark [1] particularly suitable for developing numerous applications to meet Big Data criteria 5Vs (i.e. volume, velocity, veracity, value and variety [2]). However, existing solutions with MapReduce/Hadoop based approach cannot meet some vital requirements like scalability, security and large real-time streaming [2, 3], because of some issues with I/O cost, algorithmic complexity, low-latency streaming jobs and fully disk based operation.

**2. Motivations and Research Questions**

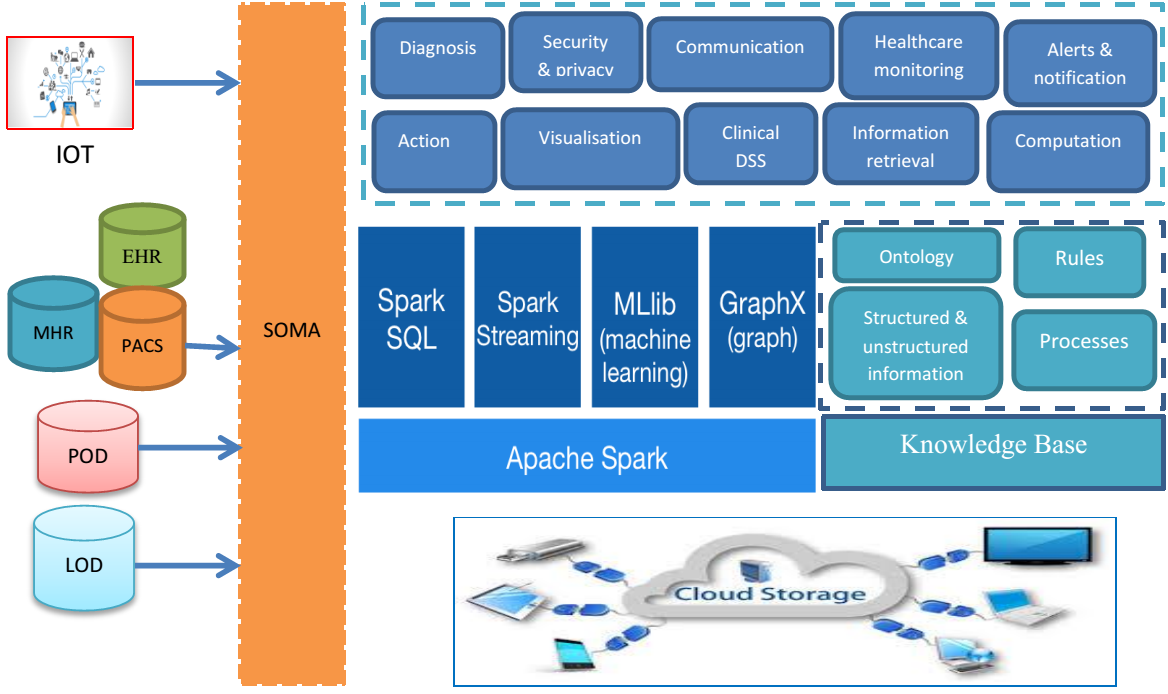
Apache Spark can be used to overcome above issues by making a difference for the healthcare applications like remote health monitoring, diseases diagnosis aids, drug references, information retrieval, clinical decision support system, patient similarity searching, prediction and visualization of the results for personalized medicine in the public health surveillance. In addition, large scale classification, clustering, predictive modeling and association rule mining can also be done with Spark’s machine learning library. Real-time streaming data from wearable sensors and IoT enabled devices can be handled and processed with the Spark’s streaming library very efficiently [1]. Besides, Hospital Information Systems (HIS) also need to deploy security and privacy of individual patient records to gain trustworthiness. Most importantly, Spark’s computing infrastructure will be suitable for distributed and real-time healthcare analytics for reducing huge healthcare costs. Considering these exciting features of Spark and limitations of existing approaches, this paper proposes a scalable, secure and real-time healthcare analytics framework to disseminate these issues.

**3. Proposed Framework**

Clinical data sources in HIS (i.e. EHRs, EMRs, PACS, wearable devices, mobile devices, IoT enabled medical devices, public open healthcare data and LOD) can be of huge importance. Spark streaming is used to handle massive streaming data coming from streaming sources. On the other hand, SparkML is used to handle big static datasets coming from static data sources. Proposed Knowledge bases with Ontology Web Language (OWL2) and Open Biomedical Ontology (OBO) are used to generate rules, ontologies and semantic annotation. Integration of heterogeneous data sources and knowledge base provides meaningful insights in this case. Security and privacy of data records are ensured with anomaly detection and privacy preserving data mining [8]. SparkML is used for recommendation and diagnosis based on structured and unstructured knowledge processing. Large scale classifications are done using Spark’s Support Vector Machine, Random forests and K-Means clustering libraries. Besides, computations, alerts, notification and healthcare monitoring are done through SparkML. Information retrieval is done through the Spark SQL and SPARQL, where Spark GraphX works as the visual analytics of the queried data.

**4. Implementation**

The *Casandra* [6] works as the cloud storage platform to ensure scalable computing over the SOMA [4] platform. SOMA and Spark streaming are used to fetch massive and heterogeneous streaming/static data in a scalable way with cluster manager Apache Mesos [7]. Streams are received as batches of data and our proposed algorithms and spark engine [1] processes these data into healthcare analytics.



analytics

Figure 1: Overview of the proposed framework

**References**

# [1].Apache Spark,<http://spark.apache.org/>, accessed 9.10.2015.

# [2]. Nambiar et al., A Look at Challenges & Opportunities of Big Data Analytics in Healthcare , 2013 IEEE Intl. Conf. on Big Data. [3]. Marco et al., “Big Data, Big Knowledge: Big Data for Personalized Healthcare”, IEEE Journal of Biomedical and Health Informatics, V-19, NO. 4, JULY 2015.

# [4] <https://www.insight-centre.org/content/soma-linked-data-platform>

[5]. Matthew H. et al. , A review of data mining using big data in health informatics, Journal of Big Data, 2014.

[6]. Apache Casandra, <http://cassandra.apache.org/>, accessed 20.8.15

[7] Apache Mesos, <http://mesos.apache.org/>, accessed 15.7.2015

[8]. Karim et al., Privacy Preserving Mining Maximal Frequent Patterns in Transactional Databases, Springer-LNCS, DOI; 10.1007/978-3-642-29038-1\_23