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

Centre for Data Analytics
Insight

Md. Rezaul Karim, Ratnesh Sahay and Dietrich Rebholz-Schuhmann

Introduction

- Healthcare devices (wearable sensors, IoT devices, EHRs), public open health data (i.e. LOD) & knowledge from biomedical literatures => emerging Big healthcare Data
- Existing healthcare analytics approaches cannot meet => scalability, security and large scale real-time streaming job efficiently
- Ex: Hadoop/MapReduce has issues => I/O cost, algorithmic complexity, low latency streaming jobs with fully disk-based operation
- Apache Spark is suitable for healthcare applications to meet Big Data criteria => 5Vs

Motivations & Research Questions

- How to handle/process real-time healthcare streaming from wearable sensors, IoT devices and static EHRs data efficiently and ease?
- Large scale classification, clustering, predictive modeling and similarity searching => key challenges for Big Data healthcare analytics
- HISs also need security & privacy of individual patients records =>Trustworthiness
- Most importantly, a robust, scalable, and real-time system needs to be deployed to reduce huge healthcare costs

Goals

Research and implementation of Spark based Big Data healthcare analytics framework for:

- Handling streaming/static healthcare data
- Remote healthcare monitoring
- Disease diagnosis aids
- Drug references
- Patient similarity searching
- Clinical DSS
- Prediction and visualization

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

Proposed Framework

❑ Massive streaming/static data are fetched with SOMA/SparkSL/SparkML/Kafka

❑ Streams are received as batches => our proposed algorithms with spark engine process batches of data into analytics for:

- Healthcare monitoring, patient similarity search, diagnosis aids, drug references etc.
- Security & privacy => anomaly detection & PPDM
- Information retrieval =>SPRQL/Spark SQL
- Visual analytics => Spark GraphX/D3

❑ Generating knowledge using rules, ontologies (OWL2/OBO) => Knowledge base

❑ Apache Mesos => Cluster manager

❑ Cloud storage platform to ensure scalable computing over the SOMA => Casandra

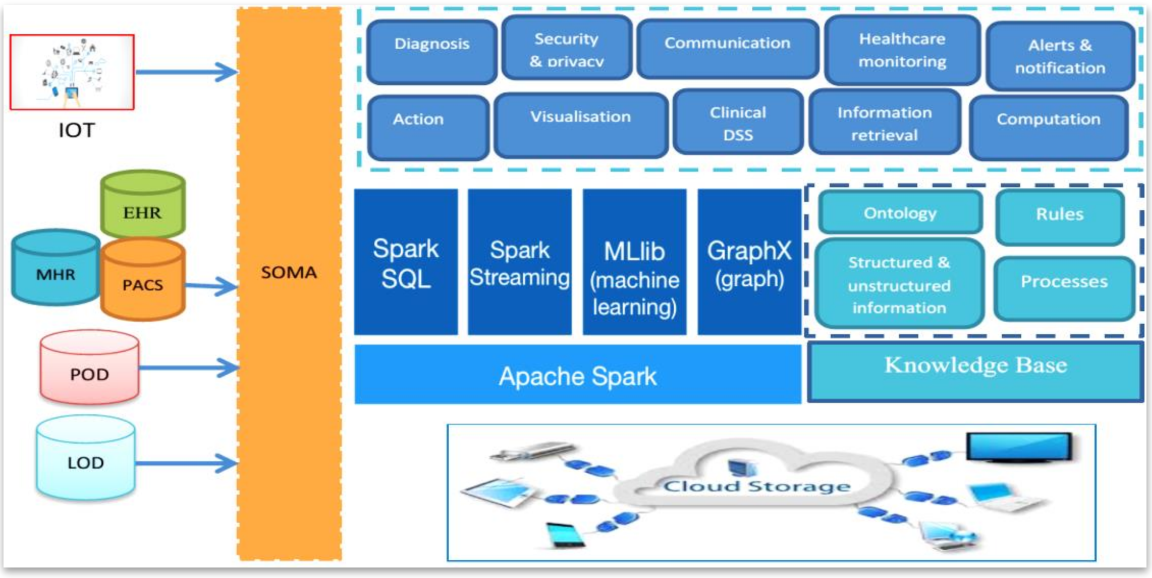

Fig-1: Conceptual diagram of the proposed framework

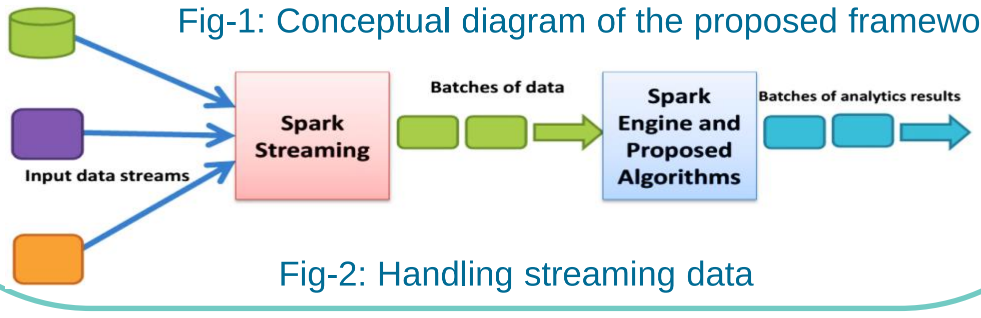
Input data streams
Spark Streaming
Batches of data
Spark Engine and Proposed Algorithms
Batches of analytics results

Fig-2: Handling streaming data

Implementation Plan

- Review & analysis of existing approaches √
- Modeling & conceptual architecture design √
- Prototyping & implementation
- Performance evaluation
- Optimization

DCU
NUI Galway OÉ Gaillimh
UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh
UCD Dublin
sfi Science Foundation Ireland