A Scalable Data Integration and Analysis Architecture for Sensor Data of Pediatric Asthma

Dimitris Stripelis[*], José Luis Ambite[*], Yao-Yi Chiang[†], Sandrah P. Eckel[‡], and Rima Habre[‡]
[*]Information Sciences Institute, [†]Spatial Sciences Institute, [‡]Department of Preventive Medicine
University of Southern California
{stripeli, ambite}@isi.edu, {yaoyic, eckel, habre}@usc.edu

***Abstract*—According to the Centers for Disease Control, in the United States there are 6.8 million children living with asthma. Despite the importance of the disease, the available prognostic tools are not sufficient for biomedical researchers to thoroughly investigate the potential risks of the disease at scale. To overcome these challenges we present a big data integration and analysis infrastructure developed by our Data and Software Coordination and Integration Center (DSCIC) of the NIBIB-funded Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) program. Our goal is to help biomedical researchers to efficiently predict and prevent asthma attacks. The PRISMS-DSCIC is responsible for collecting, integrating, storing, and analyzing real-time environmental, physiological and behavioral data obtained from heterogeneous sensor and traditional data sources. Our architecture is based on the Apache Kafka, Spark and Hadoop frameworks and PostgreSQL DBMS. A main contribution of this work is extending the Spark framework with a mediation layer, based on logical schema mappings and query rewriting, to facilitate data analysis over a consistent harmonized schema. The system provides both batch and stream analytic capabilities over the massive data generated by wearable and fixed sensors.**

Demo Video: https://www.youtube.com/watch?v=6ntm4C29L-I

I. Introduction

Asthma is the most common serious chronic disease in infants and children, affecting 9.3 percent of the American pediatric population. The Federal Interagency Forum on Child and Family Statistics estimated that in 2010, 3 out of 5 asthmatic children had at least one asthma attack during a 12 month period. However, current tools are limited in the types of information they can gather, which limits their predictive power. To improve this state of affairs, the NIH-NIBIB has established the Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) program to integrate environmental, physiological, and behavioral factors in epidemiological studies of asthma. Within this context we, at the PRISMS-DSCIC, are developing a general data integration and analysis system that enables biomedical researchers to investigate prediction algorithms for asthma attacks and eventually provide close-loop interventions.

Our system builds upon Apache Kafka and Apache Spark, which are used to integrate both sensor and traditional data sources, and to provide analytics at scale. A novel feature of our architecture is a Mediation layer, using schema mappings and query rewriting [2], [4], which maps the schemas from the heterogeneous data sources into a common harmonized schema for the analytics components to operate on.

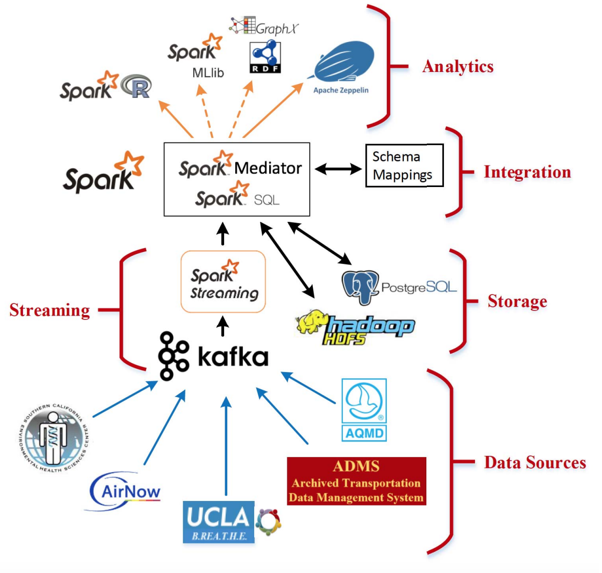

Figure 1: PRISMS-DSCIC Architecture

II. System Architecture

In this section we provide a brief description of the PRISMS-DSCIC integration and analysis infrastructure. Figure 1 illustrates the main components of the system, which we describe next in the direction of data processing, from sources to analytics. The ***Data Sources*** layer shows current sensors and web services that provide data to the system, including:

- *Southern California Environmental Health Sciences Center (SCEHSC)* provides environmental exposures based on moving sensors for selected subjects in their cohort.
- *Airnow* provides hourly measurements of ground-level ozone (O3), particulate matter (PM2.5, PM10), and overall air quality index for zip codes in Los Angeles County.
- The *LA-PRISMS Biomedical Real-Time Health Evaluation (B.REA.T.H.E)* platform provides environmental and physiological data generated by fixed and wearable sensors that measure motion activity (accelerometer and gyroscope), heart-rate, air pollution (PM, dust). Update frequencies range from 5 seconds to once per day.
- *Archived Transportation Data Management System (ADMS)*[3]) tracks traffic volume in freeways and major streets in the LA metropolitan area. For the PRISMS-

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.198

DSCIC, they provide a web service with temporally and spatially aggregated vehicle counts. We query this service every 20 minutes over a 604 grid cells of 1-mile radius, covering all of LA county.

- *South Coast Air Quality Management District (AQMD)* provides daily measurements of ozone (O3), particulate matter (PM2.5, PM10), nitrogen dioxide (NO2) and carbon monoxide (CO), as well as air quality forecasts for the Southern California region.

The ***Streaming*** layer receives data from streaming sensors using Apache Kafka. Each sensor is associated with a single Kafka Producer which writes the incoming data streams to a designated Kafka Topic for this source. The Spark Streaming component in turn invokes the Spark Kafka Consumers which pull the data out of the Kafka Topics.

The ***Integration*** layer maps the data from the different sources into a common schema using logical schema mappings and query rewriting [2], [4]. The Apache Spark SQL engine [1] handles the distribution of the queries and pushes raw sensor data and harmonized data to the Storage layer.

The ***Storage*** layer manages permanent data storage. A goal of the PRISMS-DSCIC is to build datasets for future studies of pediatric asthma. To this end, we store raw and harmonized data in HDFS and also in PostgreSQL. We plan to use HDFS storage for offline training of statistical models, and PostgreSQL for online data exploration.

The ***Analytics*** layer builds upon SparkR and Spark MLlib libraries to enable researchers to analyze real-time and historical data, e.g., to build prediction models for asthma attacks. Currently, we provide a RStudio/Rserver and Apache Zeppelin interfaces to connect to the Spark cluster, so that researchers can explore different statistical/machine learning models.

III. Demonstration

We demonstrate a spatio-temporal statistical modeling use case for heart-rate prediction based on environmental features. The analysis was performed over the SparkR framework and involves real-time collected data from the SCEHSC data source. A volunteer wore several environmental exposure sensors and a heart-rate monitor over a 24 hour period. Figure 2 shows the volunteer's trajectory from USC Health Sciences campus to Redondo Beach (morning and evening commute) and the recorded heart rate (bpm) along her track.

The main goal of this sample analysis is to identify the best window aggregation period and most informative features for heart-rate prediction, considering the continuous variability in the environmental exposure metrics. The computation involved in this use case is similar to identifying the asthma exacerbation triggers. The features are the Lung Deposited Surface Area (LDSA) concentration of airborne particles, Black Carbon (BC) exposure (related to traffic pollution), Particulate Matter (PM2.5) measured over more and less precise sensors (RHCorr_Neph and Ac1_PM), and the sound level exposure from a high and a low quality sensor (As_SoundLeveldb and Ac1_SoundLevel). We consider 4 time windows: 10 min, 30 min, 60 min and 120 min. The Figure 3 shows the ranking of the relative influence of the above features in the heart-rate prediction task, using a Generalized Boosted Model (R package gbm). Particulate matter and black carbon aggregated in 30min-1hour windows were the most predictive. Further analysis restricting the features to the seven most informative showed similar prediction accuracy as the complete set.

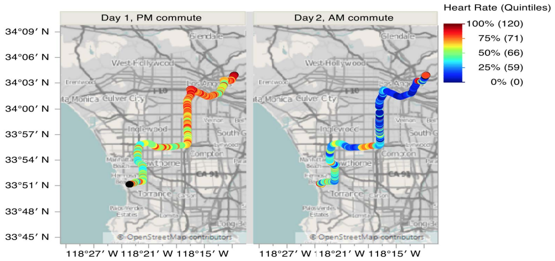

Figure 2: Subject trajectory and heart-rate

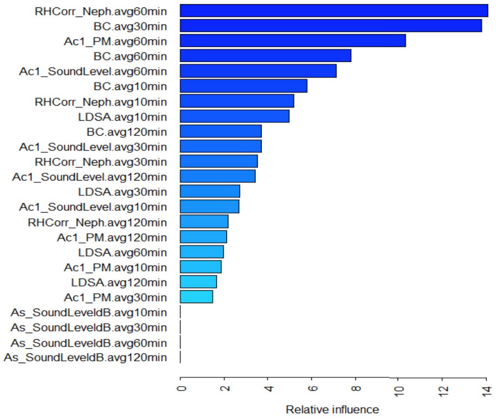

Figure 3: Exposures and window size influence

IV. Discussion

This demo shows an integration and analysis architecture for streaming sensors and traditional data sources, including a mediation layer so that analytics can be performed over a common harmonized schema. The system allows researchers to train and apply statistical models over both streaming and historical data. The application domain explores environmental, physiological, and behavioral factors for epidemiological studies of pediatric asthma.

Acknowledgment

This work was supported by NIH grant 1U24EB021996-01.

References

[1] Michael Armbrust, Reynold S. Xin, et al. *Spark SQL: Relational Data Processing in Spark*. SIGMOD, Melbourne, Australia, 2015.

[2] Alon Y. Halevy. *Answering queries using views: A survey* The VLDB Journal 10(4), 2001.

[3] H.V. Jagadish, Johannes Gehrke, et al. *Big Data and Its Technical Challenges*, Communications of the ACM, 57(7), 2014.

[4] George Konstantinidis and José Luis Ambite. *Scalable query rewriting: a graph-based approach.* SIGMOD, Athens, Greece, 2011.