MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220086

OpenMRS Analytics Engine: A FHIR Based Approach

Allan Kimaina[1,2], ScM; Jonathan Dick[1,2], MD; Bashir Sadjad[3], PhD

[1]Regenstrief Institute, Indianapolis, IN; [2]AMPATH, Eldoret, Uasin Gishu, Kenya; [3]Google Health, Waterloo, Ontario, Canada

Abstract

As the Electronic Health Record (EHR) data keeps growing in volume at an unprecedented rate, there is an increasing need for a more collaborative and scalable approach for designing and engineering clinical data pipelines. To address these two critical needs, we present a scalable analytics pipeline architecture, designed from the bottom-up to harness the power of FHIR (Fast Healthcare Interoperability Resources) for improving collaborative efforts in health data analytics and indicator reporting.

Keywords

FHIR, Health data science, Analytics-pipeline

Introduction

There is no doubt that the design and technologies used in most EHRs can barely support and sustain scalable data analytics, particularly in deployments with ever-increasing massive amounts of health data[1]. As such, we are solving challenges inherent in extracting and transforming data from EHR for analytics purposes. Some of the problems underscored by OpenMRS implementers in one of our surveys include: lack of a standard approach for provisioning and transforming EHR data, making it difficult for open-source developers and implementers to collaborate; difficulty in transforming data to reportable indicators/analysis due to overly complex EHR data models; difficulty in maintaining the current manual Extract Transform Load (ETL) pipelines due to overreliance on legacy technologies not meant for analytics; limitations of existing ETL approaches to scale workloads with the increased influx of "big data"; risk of impact on operational EHR performance while performing analytics and reporting and query data quickly. Additionally, we will demonstrate how unified analytics through standardized schema like FHIR representations could improve collaborations (e.g., indicator definitions), something which has previously been viewed as intractable.

Methods

To vastly improve the ability to generate reports and rapid analytics, we designed and implemented a robust and scalable data extraction, provisioning, and processing pipeline, engineered from the bottom-up for performance using lambda architecture. Lambda architecture encompasses two components, i.e., bulk (batch) mode and continuous (streaming) mode[2]. In streaming mode, we have created a mechanism of continuously translating new changes/events in the EHR system into FHIR resources and uploading them to the target data warehouse. On the other hand, the bulk mode will be used to efficiently extract the entire content of the EHR data, transform it into FHIR bundles, and finally write it to the target data warehouse. To perform fast indicator calculations, we utilized open-source technologies such as Apache Spark, Apache Beam, Apache Parquet, Debezium, and Bunsen. The design of the pipeline is presented in Figure 1.

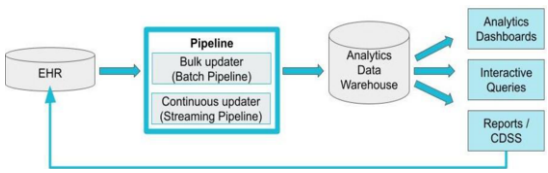

Figure 1–Analytics pipeline architecture

Results

We evaluated the pipeline performance using AMPATH's Medical Recording system, a point of care system built on top of the OpenMRS stack. AMPATH's EHR is one of the largest health record databases in sub-Saharan Africa and currently hosts over 300 million clinical observations, with over 2000 new clinical encounters being recorded every day in real-time across six counties in Western Kenya. Our preliminary analysis involved monitoring and recording the time it takes to extract patient FHIR resources for over 1,072,569 patients. To protect patient information, one authorized personnel from the AMPATH data management team was given temporary access to run the test within AMPATH's data center. Using a fairly low-end testing server (4 cores and 30 GB RAM), the batch pipeline was able to extract the entire patient resource within 2 hours 23 minutes.

Discussion

While data warehouse schema modeling is essential for performance considerations, we chose FHIR as opposed to a custom-designed schema or other standardized schemas such as Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). The primary motivation for using FHIR was that most EHR systems already speak FHIR. Additionally, using a standardized schema will drastically improve collaboration in tooling and indicator definitions. Secondly, it is critical to have the flexibility to accommodate and process the ever-increasing amount of data in the EHR in a much more efficient and scalable way. In our approach, we leveraged existing big data and stream processing frameworks while ensuring different deployment scenarios such as cloud and local setups. Lastly, due to the sensitivity of patient health records, we recognize the need for integrating a de-identification mechanism in the analytics pipeline and data warehouse - something we are yet to explore.

Conclusions

For healthcare providers aspiring to achieve end-to-end analytics, high-throughput data engineering tools and methodologies can offer an escape from many of the challenges and limitations of stretching production EHR systems to perform analytics. In this exposition, we have demonstrated how to extract data in a much more efficient way; also, to encourage collaboration between organizations, we have demonstrated how to provision these data elements in a widely known standardized schema such as FHIR; lastly, we have demonstrated how to generate sample indicators using an extremely performant approach that can scale linearly depending on server or cluster specifications.

Acknowledgements

We would like to express our heartfelt appreciation to OpenMRS, Google, and AMPATH for their incredible support.

References

[1] Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. -Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Trans Biomed Eng. 2017 Feb;64(2):263-273. doi:10.1109/TBME.2016.2573285. Epub 2016 Oct 10. PMID: 27740470; PMCID: PMC5859562.

[2] Pal G, Li G, Atkinson K. Multi-Agent Big-Data Lambda Architecture Model for E-Commerce Analytics. Data. 2018; 3(4):58

Address for correspondence

Allan Kimaina : akendagor@ampath.or.ke
Jonathan Dick: jdick@ampath.or.ke
Bashir Sadjad: bashir@google.com