Towards a heterogeneous, polystore-like data architecture for the US Department of Veteran Affairs (VA) enterprise analytics

Conference Paper · December 2016

DOI: 10.1109/BigData.2016.7840896

CITATIONS
12

READS
170

3 authors, including:

Edmon Begoli
Oak Ridge National Laboratory
60 PUBLICATIONS **815** CITATIONS

SEE PROFILE

Towards a Heterogeneous, *Polystore*-like Data Architecture for the US Department of Veteran Affairs (VA) Enterprise Analytics

Edmon Begoli
Derek Kistler
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
1 Bethel Valley Rd.
Oak Ridge, Tennessee 37831, USA
begolie@ornl.gov, kistlerde@ornl.gov

Jack Bates
Business Intelligence Service Line
US Department of Veterans Affairs (VA)
810 Vermont Ave. NW
Washington, DC 20420, USA
jack.bates@va.gov

***Abstract*—The *Polystore* architecture revisits the federated approach to access and querying the standalone, independent databases in the uniform and optimized fashion, but this time in the context of heterogeneous data and specialized analyses. In light of this architectural philosophy, and in the light of the major data architecture development efforts at the US Department of Veterans Administration (VA), we discuss the need for the heterogeneous data store consisting of large relational data warehouse, an image and text datastore, and a peta-scale genomic repository. The VA's heterogeneous datastore would, to a larger or smaller degree, follow the architectural blueprint proposed by the polystore architecture. To this end, we discuss the current state of the data architecture at VA, architectural alternatives for development of the heterogeneous datastore, some relevant use cases, the anticipated challenges, and the drawbacks and benefits of adopting the polystore architecture.**

I. Introduction

The United States Department of Veterans Affairs (VA) is a US federal government-run benefit system that provides health and other benefits to US military veterans. VA employs nearly 345,000 people at hundreds of Veterans Affairs medical facilities, clinics, and benefits offices, and is responsible for administering benefits programs for 21.7 million veterans, their families, and their survivors. The Veterans Health Administration (VHA), an organization within the VA system and the largest healthcare system in US, is responsible for providing all forms of health care to veterans, running community based health centers and regional medical centers, and for conducting biomedical research. As of 2016, VHA supports 8.97 million veterans [1]. As part of this mission, VA manages large quantities of data related to the veterans' population, including data related to the clinical services, treatments, and general healthcare. As part of the overall healthcare delivery system, VA is increasingly looking into advanced analytics methods to improve the quality of care, understand the underlying factors related to medical conditions impacting its patient population, and to improve the services and delivery of care for the veterans' population. For this to happen, VA needs data assets and systems that are consistent, timely updated, flexible, and easily accessible for the applications of comprehensive and diverse types of analyses. The efforts described in this paper, which are part of the overall CHAMPION program [2], are aimed at examining the data systems architectures that would support the efficient data management of the massive and heterogenous new data sources, and the implementation of the comprehensive analytics as outlined in the previous paragraph.

II. Information Systems and Datastores at VA

VA is a large enterprise, covering a broad group of services and administration of benefits beyond healthcare services. In this paper, we focus on the large data and analytics systems in support of the delivery of healthcare services and medical research, and these are discussed in the following sections.

A. VA's EHR - VistA

VistA – Veterans Health Information Systems and Technology Architecture – is an integrated Electronic Health Record (EHR) information technology system with application packages that share a common data store and common internal services. The data store and VistA kernel are implemented in the MUMPS (or M) computer language, and the Computerized Patient Record System (CPRS) graphical user interface (GUI) is implemented in Delphi. VistA is deployed universally across VHA at more than 1,500 sites of care, including each Veterans Affairs Medical Center (VAMC), Community Based Outpatient Clinic (CBOC), and Community Living Center (CLC), as well as at nearly 300 VA Vet Centers.

B. Corporate Data Warehouse (CDW)

Some data from VistA is accessed directly from the local system for operational reporting, but the majority of the data is streamed into the regional data warehouses from where it is further aggregated into the central warehouse called Corporate Data Warehouse (CDW). CDW stores data that is a direct representation of the medical records as maintained in VistA

as well as other data relating to the clinical and administrative aspects of healthcare. As such, some data is discrete, tabular, and some is unstructured and textual, given that physicians store their notes in the medical records. The same applies to pathology reports and radiology notes. While this is highly useful data, it is more difficult to access, use, and process relationaly than through a specialized text analytic system.

C. VA Informatics and Computing Infrastructure (VINCI)

The mission of VINCI is to improve the healthcare of Veterans by providing researchers access to integrated national datasets and tools for analysis in a secure, high-performance computing environment. The objective of the VINCI program is to provide a secure, high-performance computing environment for researchers to access data, to construct integrated databases from national clinical and administrative data sets, to consolidate processes and procedures for access to data ensuring compliance with VA policies, to provide tools to analyze data, report results, support informatics research, and to reach out to the research community.

D. Million Veterans Program (MVP) Genomic Database

The Department of Veterans Affairs Million Veteran Program (MVP) [3] is a national, voluntary research program that covers sequencing, storage, and management of millions of US veterans' genomes. At the time of the writing of this paper, 500,000 veterans have volunteered their sequences, making MVP genome database, as of August 1st, 2016, the largest genomic database in the world [4]. The data stored in MVP is initially being used to address gaps in the genetics of a disease/trait of interest, but the other areas such as the integration with the CDW data, phenotyping, statistical genetics and others will be considered. At this time, the process of accessing and querying the data is manual. Researchers need to request access to the data, and upon approval they are given data extract to work on an HPC (High Performance Computing) [5] cluster. The ultimate goal is to make this data available in an entirely automated fashion, and available to other analytic and database engines that need it, and follow the required security policies.

III. Analytic Uses of Data

The primary driver for having the data warehouse at VA is for having the strategic decision support, data analysis, and business intelligence capability. In the context of VA, the strategic analytic domains are operational and clinical analysis, business intelligence, and medical research through the VINCI and other research programs.

Currently, CDW [6] is used to support analysis and operational reporting aimed at the improvement of health care quality, system-wide patient safety, support of the medical and rehabilitation research, optimization of the service delivery and efficiency, studies in the improvement of the patient population satisfaction, and for coordination and information sharing with other agencies related to coordination of emergency response, participation in biosurveillance activities, and simulations.

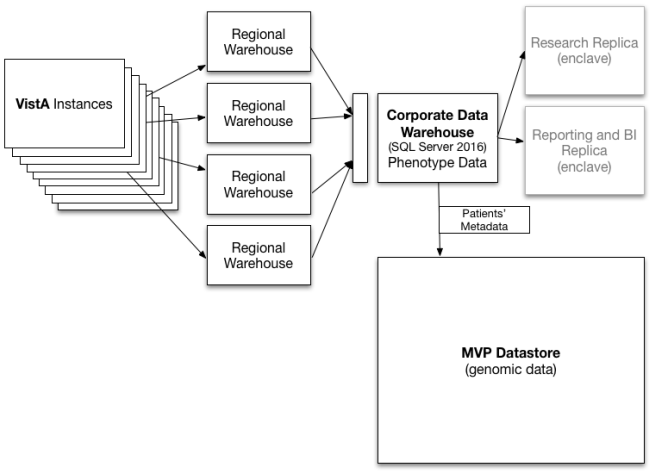

Fig. 1. The current state of the data systems used for analytics at VA.

A. Future Uses of Data

Medical data is inherently heterogenous. In addition to medical records, which contain structured, discrete, and unstructured textual notes (physician notes, etc.), clinical data often includes imaging results (MRI, PET, CT, etc.) and increasingly genomic profile of the patients. Unstructured data, such as physician notes, pathology, or radiology reports, is essential for complete understanding of the patients medical history and clinical conditions. With the MVP program and the establishment of the large repository of genomic data, the data landscape at the VA is only going to get more complicated, and the analytic challenges will accordingly magnify. These new data sets are going to offer opportunities for life-changing improvements and scientific breakthroughs. As part of the overall evolution of the analytic capabilities, VA will seek to improve or develop its analytic capabilities in the areas of operational and clinical analytics, business intelligence and reporting, predictive analysis and optimization, cohort studies [7][pp.1-405], disease studies, population health [8], and personalized and genomic medicine.

Interesting areas at the intersection of genomics and cohort studies [3] that may use such a large well-characterized cohort include: investigations of gene-environment (lifestyle) interactions, genome-wide associations, pharmacogenomics, and nutrigenomics. Ultimately, the hope is that the findings in these areas will lead to better ways to prevent, detect, and treat disease.

B. The Case Studies of Interest

Some of the research priorities for Veterans Administration are studies in the causes of the prostate and lung cancer, and Post Traumatic Stress Disorder (PTSD).

With the access to the heterogeneous data outlined above, and with the implementation of the polystore-like architecture, we would like to, among other phenomena, explore:

- correlations between the patient's medical history, phenotype and genotype profile (e.g. presence of the prostate

cancer specific mutations [9], [10]) as they relate to the prostate cancer,
- impacts and correlations between different diagnostic procedures, treatments, and the outcomes based on the patients genetic profiles, and
- correspondence between the patient's phenotype, information recorded in the physician notes, and the genetic profile [11], and their potential effects on the occurrence, progression, and the effectiveness of the treatments for the PTSD.

IV. Towards a Heterogeneous Data Architecture

A VA's diverse data is stored in the standalone, individual systems that best reflect the structure of the particular datasets, and that provide the most efficient means for data management and access. Modern medicine, however, requires looking into all sources of information to form a complete and well-informed picture about the patient and the diagnosis. The integrative approach becomes even more important at the time of precision medicine. David et al. state as one of the challenges for implementation of genomic medicine the need to integrate the aspects of the genomic datasets with clinical data ([12], Table 2). We see a polystore approach, as presented by Duggan et al. [13], as a promising alternative to the problem of heterogeneous data integration for the sake of comprehensive, timely, and integrated analysis. To make the polystore approach truly useful and validated, it has to be examined against the alternatives and anticipated challenges (described in the next two sections), and practically evaluated for performance, usability, and reliability.

A. Alternatives Under Consideration

Re-architecting the VA data enterprise is a challening undertaking. First, we need to take into consideration analytic clients that use existing data systems. Second, we need to account for the future growth and the evolution of the current analytic use cases. Finally, we need to account for the new, precision medicine use cases, and specifically account for the growth and incorporation of the MVP dataset. In the following sections, we present main architectural alternatives under consideration, and we discuss some preliminary observations about the benefits and drawbacks of each of these alternatives.

1) MPP-based Architecture: Currently, SQL Server 2016-based CDW is the foundation of the VA's enterprise analytics. To support the future analytic growth, the natural choice is to scale the existing functionality and capacity of CDW and to migrate the single processing database architecture onto a massively parallel processing (MPP) database platform. This migration would add significant capacity and throughput to the CDW architecture, while continuing to make the same data available to the same clients. To address the unstructured data analysis needs, some data (unstructured, or mixed structured and unstructured) could be made available to researchers through a dedicated "discovery"/research sandboxes, similar to how it is done today with VINCI. The genomic, MVP data could remain stored in a separate MVP repository, and the segments of that would be indexed, with the pointers to it stored in the CDW MPP for easier recognition and retrieval.

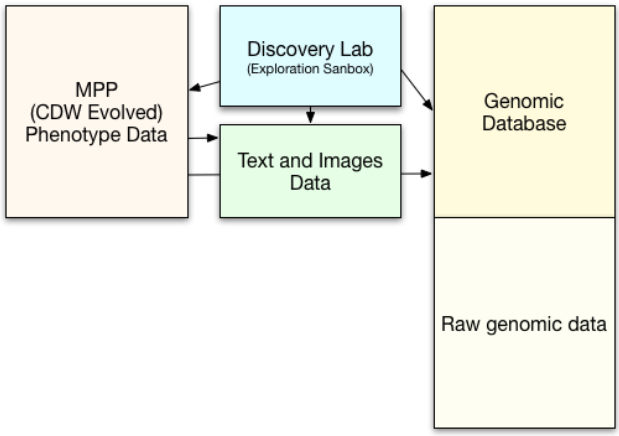

Fig. 2. An MPP-dominated alternative for future VA data architecture.

The benefits of this architecture would be mostly for the VA's data warehousing and business intelligence community. The existing asset (CDW) would receive a boost. The user community needing datasets that are beyond what is in CDW would have access to both genomic, structured, and unstructured data through a "sandbox". Genomic data would remain where it is, and it would have to be separately managed and scaled, but important segments of it would be indexed and accessible through an MPP. The drawbacks of this architecture are multiple. Its first, hypothetical drawback is in its partially manual and partially inconsistent nature of data management. The enterprise data managed this way will likely never be entirely uniformly and centrally accessible. Consequently, the access to it from the clients and through the queries would not be consistently optimizable.

2) A "Big Data" Solution: We call this a "Big Data" solution as Apache Hadoop and Spark have virtually become synonyms for the "Big Data" term. With this solution, all of the enterprise data would be migrated onto a parallel file system such as HDFS [14] and then accessed from there using some of the available tools from the Hadoop and Spark-compatible universe.

There are several drawbacks of this approach. First, the canonical way to implement this architecture is to put as much of the data, including both genomic and non-genomic data, on HDFS, because most of the tools and APIs in this technology set work the best and are the best supported with HDFS [15]. While this is the best understood way for dealing with the data with Hadoop and Spark, we do not have sufficient evidence that HDFS is a performance-wise optimal medium for storing and accessing both structured and unstructured data at this scale or processing it via tools such as ADAM [16]. Second, this approach will still likely require manual interventions and substantial engineering effort to reconcile the MPP, genomic and unstructured data stores, and create a uniformly accessible

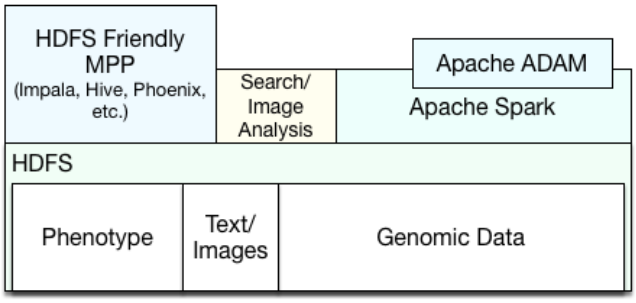

Fig. 3. A "Big Data" alternative.

and consistent API layer.

3) Polystore Approach: Given the considered alternatives, polystore architecture emerges as the logical medium between the two alternatives. With polystore, such as BigDAWG [13], the data remains in the native systems that are optimized for the processing of the respective data types and structures, but the query federation layer of the polystore takes care of the query processing and optimization in a uniform manner.

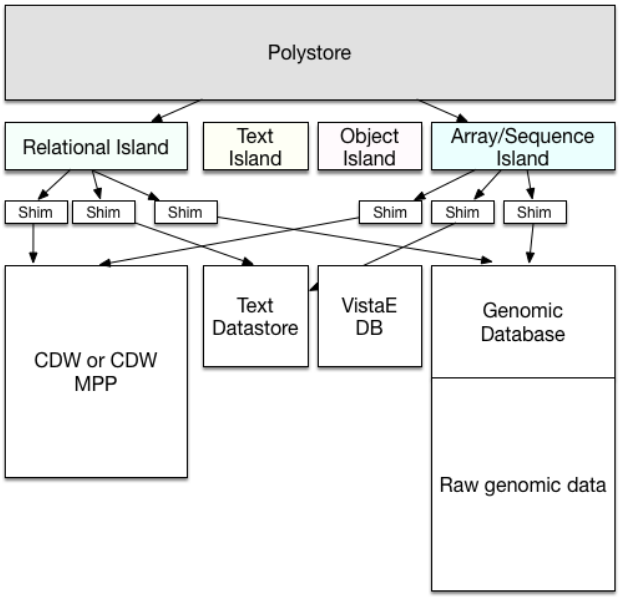

Fig. 4. A *polystore* alternative for the future architecture.

The polystore architecture is a research proposition, so the comparison between the MPP-dominated architecture, which has decades of the maturity, and HDFS/Spark/ADAM-based which is also past the research state, is hypothetical and speculative at best. However, the advantages that the polystore approach is offering, are obvious – the data remains in the original hosts, there is a uniform and familiar processing API, the impacts on the existing systems are minimal, and there is a method for optimization of the federated queries across the diverse, heterogeneous datasets. The drawbacks, given the early state and immaturity of the technology cannot be as clearly stated as for the previous two alternatives.

V. Anticipated Challenges

The following challenges are, at this point, mostly speculative, and are based on our experience [17], [18] on the previous efforts in the same domain that shared many of the characteristics and that led to our deepening interest in the examination and prototyping of the polystore architecture.

A. Query Translation and Optimization

Query translation and optimization are recognized as general, open research problems in this area polystore. Other challenges related to polystore query translation and optimization might more specific to the VA's scenario. The problems we anticipate are:

- a 2-3 orders of magnitude imbalance in the size (and structure) of the phenotype data, unstructured data, and genomic data;
- nascent understanding of the nature of the genomic/life-sciences queries in relation to the VA's healthcare services and research mission; and
- equally nascent understanding of the expression requirements and the structure of the query language for the VA's heterogeneous dataset.

B. Security and Preservation of Privacy

Healthcare data is heavily governed by the US federal standards for patients' privacy protection (HIPAA [19], specifically) which requires implementation of specific controls and rules at both the client and the database level. Moreover, HIPAA controls typically translate to policies that govern access rights to the particular database. In the case of VA, some instances of CDW are designated for researchers; some are designated for clinicians; some are for operations staff. Genomic store at this point is entirely designated for research use, and is limited in both technical aspects of access and policies to manually approved group of users. The polystore implementation in our domain will need to have architectural mechanisms to address this requirement.

C. Analytic APIs for Imagery, Textual and Genomic Data

Just as the data storage and database management technologies cannot fit any more in the same mold [20], so the front-end application programming interfaces (APIs) will likely need to be "polyglottic" - offering access to the data and analytic functions in multiple programming languages, or to offer extensions to the standard query languages that would be easy for business analysts to use, and expressive enough for the data scientists to work with them as well. Analyzing imagery, text or genomic data is a very specialized task, performed by people with deep backgrounds in these fields. Thus, the API will either need to be developed for these specialists (e.g. image processing, text topic models, etc.), or a simplified, uniform API will need to be developed where the underlying specifics would be rolled up to a simple API

without losing functionality. The middle ground could be User Defined Functions (UDFs).

1) Multilayer Mapping to National Standards (e.g.OMOP): OMOP [21] is an emerging common data model and vocabulary standard for representation of medical data. It is a logical model with physical, relational implementation. At the time of the writing of this paper, it is in version 5, and rapidly emerging, and we play a role in the emergence of this standard. We are currently challenged to understand how to map other domains of the medical sciences.

VI. Future Work

As part of our current infrastructure engineering, we will be researching, prototyping, and deploying very large raw file storage, and related networking solutions, and we expect that these findings will be informative for the performance baselining for the *VA Polystore* prototype.

Furthermore, once the infrastructure is in place, and the data is deployed in its respective repositories (CDW ORNL enclave, genomic MVP repository, etc.), we will look to deploy the current version of the BigDAWG along with the associated technologies (e.g. SciDB), and to develop a small prototype and a technology demonstrator of the *VA Polystore*. As part of this effort we will evaluate:

1) general usability, and tradeoffs between the flexibility and expressiveness of the approach promoted by the polystore architecture such as BigDAWG and less federated, heterogeneous data analysis approaches (e.g. Apache Spark with ADAM),
2) a programming model and implementation of VA-specific UDFs, and
3) optimizations and customizations to BigDAWG or similar architecture required to make *VA Polystore* useful in a practical setting, if possible.

We intend to publish the findings of our evaluations, customizations and the experiments.

Acknowledgments

The authors would like to thank The Department of Veterans Affairs for the continuing support of research presented in this publication under the VA-DOE CHAMPION program.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. This research has been made possible by a Lab Directed Research and Development grant at Oak Ridge National Laboratory.

References

[1] "VA Quick Facts," https://goo.gl/ZZYFm5, accessed: 2016-10-04; VA link to PDF. Shortened on Google.

[2] "Tri-Agency Partnership Working to Tailor Cancer Care Based on Genes, Proteins," http://www.va.gov/opa/pressrel/pressrelease.cfm?id=2810, accessed:2016-10-08.

[3] "MVP - For Researchers and Research Partners," http://www.research.va.gov/MVP/researchers.cfm, accessed:2016-10-04.

[4] "Million Veteran Program is now largest genomic database in the world," http://www.blogs.va.gov/VAntage/29719/million-veteran-program-now-largest-genomic-database-world/, accessed:2016-10-04.

[5] "Thinking big – Handling data from the Million Veteran Program is the health informatics challenge of an era," http://www.research.va.gov/currents/dec11-jan12/dec-jan12-01.cfm, accessed:2016-10-04.

[6] "Corporate Data Warehouse (CDW)," http://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm, accessed:2016-10-04.

[7] A. Morabia, *A history of epidemiologic methods and concepts*. Birkhäuser, 2013.

[8] D. Kindig and G. Stoddart, "What is population health?" *American Journal of Public Health*, vol. 93, no. 3, pp. 380–383, 2003.

[9] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva *et al.*, "Integrative genomic profiling of human prostate cancer," *Cancer cell*, vol. 18, no. 1, pp. 11–22, 2010.

[10] I. Agalliu, E. Karlins, E. Kwon, L. Iwasaki, A. Diamond, E. Ostrander, and J. Stanford, "Rare germline mutations in the brca2 gene are associated with early-onset prostate cancer," *British journal of cancer*, vol. 97, no. 6, pp. 826–831, 2007.

[11] M. C. Cornelis, N. R. Nugent, A. B. Amstadter, and K. C. Koenen, "Genetics of post-traumatic stress disorder: review and recommendations for genome-wide association studies," *Current psychiatry reports*, vol. 12, no. 4, pp. 313–326, 2010.

[12] S. P. David, S. G. Johnson, A. C. Berger, W. G. Feero, S. F. Terry, L. A. Green, R. L. Phillips, and G. S. Ginsburg, "Making personalized health care even more personalized: insights from activities of the IOM genomics roundtable," *The Annals of Family Medicine*, vol. 13, no. 4, pp. 373–380, 2015.

[13] J. Duggan, A. J. Elmore, M. Stonebraker, M. Balazinska, B. Howe, J. Kepner, S. Madden, D. Maier, T. Mattson, and S. Zdonik, "The BigDAWG polystore system," *ACM Sigmod Record*, vol. 44, no. 2, pp. 11–16, 2015.

[14] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. IEEE, 2010, pp. 1–10.

[15] D. Borthakur, J. Gray, J. S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash *et al.*, "Apache hadoop goes realtime at facebook," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, 2011, pp. 1071–1080.

[16] M. Massie, F. Nothaft, C. Hartl, C. Kozanitis, A. Schumacher, A. D. Joseph, and D. A. Patterson, "ADAM: Genomics formats and processing patterns for cloud scale computing," *University of California, Berkeley Technical Report, No. UCB/EECS-2013*, vol. 207, 2013.

[17] E. Begoli and J. Horey, "Design principles for effective knowledge discovery from big data," in *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 joint working IEEE/IFIP conference on*. IEEE, 2012, pp. 215–218.

[18] E. Begoli, T. Dunning, and C. Frasure, "Real-Time Discovery Services over Large, Heterogeneous and Complex Healthcare Datasets Using Schema-Less, Column-Oriented Methods," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2016, pp. 257–264.

[19] C. for Disease Control, Prevention *et al.*, "HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services," *MMWR: Morbidity and mortality weekly report*, vol. 52, no. Suppl. 1, pp. 1–17, 2003.

[20] M. Stonebraker, S. Madden, D. J. Abadi, S. Harizopoulos, N. Hachem, and P. Helland, "The end of an architectural era:(it's time for a complete rewrite)," in *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 2007, pp. 1150–1160.

[21] "Observational Medical Outcomes Partnership," http://omop.org/, accessed:2016-10-05.