Eleyon SOUTH ASIAN JOURNAL OF ENGINEERING AND TECHNOLOGY

Full Length Article

Big Data – Driving a Revolution in Personalized Medicine

S. Jafar Ali Ibrahim [a,*], Dr. M. Thangamani [a], K. Pavith Kumar [b]

[a] *Doctoral Research Fellow, Anna University, Chennai, Tamilnadu*
[a] *Assistant Professor, Kongu Engineering College, Perundurai, Tamilnadu*
[b] *UG Scholar Kongu Engineering College, Perundurai, Tamilnadu*

***Corresponding Author**
S. Jafar Ali Ibrahim
Email: jafartheni@gmail.com

ABSTRACT: Precision medicine has been thought to revolutionize how we improve health and treat disease. Today, most medical treatments are designed for the average patient using the "one-size-fits-all" approach. However, in many cases, this approach isn"t effective because treatments can be very successful for some patients but not for others. Precision medicine is a field of medicine that takes into account individual differences in people"s genes, micro biomes, environments, family history, and lifestyles to make diagnostic and therapeutic strategies precisely tailored to individual patients. Precision medicine is a newer term referring to a similar field compared to another term "personalized medicine". The term „precision medicine" entered the scientific lexicon in 2008 when business strategist Clayton Christensen, of Harvard Business School in Boston, coined the expression to describe how molecular diagnostics allows physicians to unambiguously diagnose the cause of a disease without having to rely on intuition [1-4]. The name didn"t gain traction until 2011 when a committee convened by the US National Research Council laid out a blueprint for modernizing the taxonomy of disease on the basis of molecular information such as causal genetic variants, rather than a symptom-based classification system. They called the report Toward Precision Medicine

1 Introduction

Big Data platform From an architectural perspective, the use of Hadoop in our applications as a complement to existing data systems is important: IT offers an open-source technology designed to run on large numbers of connected servers to scale-up data storage and processing in a very low cost and it is proven to scale to the needs of the largest web properties in the world. The Hadoop"s architecture offers new venues for data analysis including [5]:

1. Schema on reading: Users can store data in the HDFS (Hadoop data file systems) and then design their schema based on the requirements of the application.
2. Multi-use, Multi-workload data processing: Multiple users can have access to a shared data set at the same time for close to real-time analysis.
3. Lower cost of storage.
4. Data warehouse workload optimization.

As Apache Hadoop has become more popular and successful in its role in enterprise data architectures, the capabilities of the platform have expanded significantly in response to enterprise requirements. For example in its early days, the core components of a Hadoop system has been represented by HDFS storage and MapReduce computation system. While they are still the most important ones, many other supporting projects have been contributed to the Apache Software Foundation (ASF) by both vendors and users. These Enterprise Hadoop capabilities are aligned to the following functional areas that are a foundational requirement for any platform technology: Data Management, Data Access, Data Governance & Integration, Security and Operations [6]. The following architecture is an amalgam of Hadoop data patterns that we designed to use of Hortonworks Data Platform (HDP) in Mayo"s health care systems which are shown in Figure 1. HDP is powered by Open Source Apache Hadoop. HDP provides all of the Apache Hadoop projects necessary to integrate Hadoop as part of a Modern Data Architecture [6]. Based on our architecture, we store our datasets from different resources including EHRs, Genomics, and Medical Imaging into the Hortonwork repository and then use scripting tools like Pig and Hive to clean and prepare our data. One of the applications of an implementation of this architecture at Mayo Clinic is data retrieval and cohort creation. There are many data sources available in different departments of Mayo Clinic and each one includes millions of EHRs data and creating cohort is one of the main steps in each project. Using spreadsheets for extracting records from million records of EHR data based on the cohort criteria is a

time consuming and painful job. Pig is one of the big data tools that produce a sequence of MapReduce programs to run complex tasks comprised of multiple interrelated transformations. In one of our projects about the integration of different data sources such as lab results, medications, and patient demographics to predict survival score of each heart failure patients, our cohort is the patients with heart failure diagnosis event with at least one EF (Ejection Fraction) value within three months of the heart failure diagnosis date. To create our cohort, we need to extract our desirable records from the aggregation of four large datasets including one heart failure clinical trial and three EHR datasets from different Mayo's clinical systems. Using any spreadsheet-based tool or even SQL to retrieve data from these datasets is almost impossible. We implemented our cohort criteria in the form of pig queries in three steps: we filter all patients with heart failure ICD9 code, then in the second step, we join the results of the first query with the patients EF records, and finally the results of the second query is being filtered based on the time intervals defined in the cohort by domain experts and clinicians. Pig translates our queries to a sequence of MapReduce jobs and the jobs are sent to the servers sequentially. Using pig to create our cohorts is faster and easier than any other tools. Our dataset includes more than 150 million patient records that require usage of parallel querying and computation. To compare the performance of Pig with other tools like SQL, we ran a simple test on a data file including one million rows of data and a simple operation like AVERAGE. The SQL took 18 minutes to run but Pig based alternative ran in less than two minutes on a cluster with just two nodes [7].

Big Data driving Personalized Medicine Revolution

Researchers are poised to make huge advances in medicine, particularly in how we treat cancer and arthritis. See how big data and IT are contributing. Drugs can be expensive, difficult to research, hard to get approved and, according to a recent report, don't work on large parts of the population. These factors likely put a great deal of pressure on Pharmaceutical companies to research drugs that have the highest probability of turning a profit rather than those that could help the most people. But this paradigm may be shifting with the help of IT and big data. The industry has found new ways IT and big data are making a major impact on the way drugs are being researched by helping create more effective trials. Before we examine the benefits IT is bringing to this arena, let's try to understand what's wrong with the traditional (and ongoing) way most drugs enter the trial.

Most drugs don't work. Statistics published in the journal Nature [8] show that among the 10 highest grossing drugs prescribed in the US, even the best work in only one in four patients. Some work in only one of twenty-five. Statins, commonly prescribed cholesterol drugs, work correctly in only one in fifty patients, according to the article. Nature cites multiple reasons for this, but the basic is that our different genetic makeups (genome), proteins in our body (proteome), and body flora (the bacteria and other stuff that grows inside of us that we don't like to think about) affect how drugs work. Over the past few decades, medicine has often (and sometimes legitimately) been accused of focusing too heavily on Western patient pools (cohorts), excluding minorities and people from other countries, and therefore of doing a bad job of creating drugs that work for all ethnicities.

As drug companies try to address those criticisms and create more accurate drug trials, the cost and time of developing drugs can increase due to delays in creating patient pools, difficulty finding pools with the right genetic makeup, and the need to increase the length of trials to find drugs that will work for greater portions of the population. Add it up, and the cost to produce a new drug is now $2.6 billion [9] and rising 8.5% every year, according to the Tufts Center for the Study of Drug Development. This is an untenable situation. Changing the way we design and administer treatment trials, using big data to bring "personalized" or "precision" medicine to drug trials and research, can potentially reduce costs, allow the right drugs to be prescribed faster, and improve outcomes at lower costs. It also may mean faster drug development with easier margins. It is one of those rare win-win for pharmaceutical companies and patients. Take a look at following new ways IT and Big Data are helping drug trials that are currently going on.

Create The Precision Medicine Initiative Cohort Program

What if we could study the ongoing health records of more than 1 million people to learn which individuals respond to certain types of drugs, are at risk for a certain disease, maintain health and fitness, age, and die?

That's the goal of the $130 million[10] Precision part of a precision health initiative, with an overall $215 million annual budget) being done by the National Institute of Health in the US. The goal is to enroll over 1 million Americans in the cohort in the next three to four years. The data (anonymous, of course) from all 1 million individuals will be furnished to any interested researcher who wants to study one of the largest cohorts ever made available. Same like that we can create a cohort program with the following aspects. All members of the cohort will have their genomes sequenced, and their health

history, lifestyle habits, and environmental exposures tracked. By doing so, the study will yield a treasure trove of big data. It will allow people to track the effectiveness of medicine based on genetic markers and identify certain biomarkers that signify that people might be at risk for a given disease.

The cohort could also serve as a platform for smaller trials. For instance, if you needed 100 people with a specific genetic trait who are also taking a certain drug, it could take years to develop a cohort for a study. The Precision Medicine Initiative may be able to identify the people you need in minutes. The NIH also aims to use mobile apps and information from the trial to help the people in the trial lead healthier lives. The agency hopes that by their example they will encourage healthier lifestyles in the general population. IT advances in mobile devices, cheaper genome sequencing, databases, big data, and electronic health records make knowledge and health goals possible.

Create a Molecular Analysis model For Therapy Choice Trial

A cancer research program is intended to be a major part of the Health Department Precision Medicine Initiative [12]. It will enroll about 1,000 people in an effort to match specific types of tumors with specific medicines [13]. The program will seek out people with tumors that have failed to respond to standard cancer treatments and match those tumors with drugs known to have better outcomes based on certain genetic markers. The hope is to build a database of drugs that have positive effects on different types of tumors in order to get the best treatment to patients the fastest. Targeted treatment-based clinical trials have shown benefit in molecular subgroups of patients. Of further interest, it appears that genomics and immunotherapy are coupled to each other since the immune system recognizes neo-antigens produced by the mutanome. Hence, some of the most important markers predicting response to immunotherapy are genomic markers, PDL1 amplification (for PD-1/PD-L1 checkpoint inhibitors), and tumor mutational burden [14].

The study should be an ongoing ode to personalized medicine and big data, as it will continue to grow new "arms" to study new types of tumors and track new medicines and their effect on certain types of tumors. This has the potential to be the most important cancer trial in history, as it hopefully will unlock the secret to curing multiple rare and fatal types of cancer by matching the individual's genome to the right drug. This would not have been possible until recently, because genome sequencing would have taken too long. Faster computers, better software, and better databases are at the heart of the success of trials like these.

Wisdom Study

One problem in the treatment of cancer, especially breast cancer, is the false positive. Ten percent of the time, women who get a screening mammogram are called back for further study, but only 5% of the women who get called back actually have cancer. The Wisdom Study is designed to enroll 100,000 women to see if mammograms are really the best way to detect breast cancer. The interesting thing about the study from an IT perspective is that it isn't using an electronic health record provider or a specialized database to manage the study. It is using Salesforce, the cloud-based CRM. According to the director of the research of Salesforce organization [15-16], Salesforce offers more flexibility, better data management, and the ability to more easily treat patients like people. When something as ubiquitous as Salesforce is being used for clinical trials on a grand scale, you know IT is having a major impact. The cloud, consumerization of IT, and cloud analytics are all responsible.

Tanner Project

Another new concept in drug and health trials is the "N-of-1" concept[17]. Most trials require hundreds or thousands of "N's" (people enrolled in the study) to make sure the project has statistical significance. But we're starting to realize that, with some trials, larger cohorts mean you have X's, Y's, and Z's mixed in with your N's. In other words, on a genetic level, we're comparing apples to oranges and we're adding static to the data. The Tanner Project is specifically looking for N's-of-1, people with variations from the statistical norm. The study looks for "stage zero" signs of hereditary diseases. In other words, it is looking for signs of potential diseases before people officially have the disease. The hope is that in finding those markers before the disease begins, treatments can be started sooner. For instance, doctors often look for a specific blood protein as a sign of ovarian cancer. In most people, the count of that protein has to be 30 to 35 to be significant. In some people, a much lower number can be a sign of the beginnings of cancer. If it can be discovered why the number is lower in some people (the N-of-1), the information can be extended to the hundreds or thousands of similar traits, turning an N-of-1 into a statistically useful study and helping people with that genetic and physical makeup.

Arthritis Trials

You probably noticed that most of these trials I have mentioned so far have centered on cancer. That disease is a chosen focus of precision and personalized medicine because of the huge variation of success in cancer

treatment drugs based on genetic factors[18]. However, cancer isn't the only disease where trials are being conducted. Another common area is arthritis treatment. Multiple trials have been conducted to determine the best drugs and dosage for those suffering from the disorder. Specifically, rheumatoid arthritis, which has several different biomarkers, has been shown to have significant variance in treatment response. Ultimately, there's no reason -- with faster DNA sequencing, cloud analytics, and cheaper storage -- why large-scale precision medicine trials couldn't be used for any disease.

Conclusion

Drug trials and large-scale gene studies aren't the only ways IT is advancing personalized medicine. Simple things like the Fitbit and other consumer devices may eventually lead to breakthroughs in health based on big data. Companies such as UBC, among others, are shaving costs simply by using large databases to do a better job of finding patients to enter studies. While the NIH's Precision Medicine Cohort is open to all, many drug companies are starting similar (or larger) databases. In some cases, those databases are a result of pooling data within the pharmaceutical industry. All of this is powered by major breakthroughs in IT and analytics. If you've ever helped develop technology such as a database, cloud architecture, a supercomputer, big data, mobile, or networking, you've probably had a hand in inspiring these studies that could one day help conquer some horrible diseases. To date, precision medicine programs have accomplished many preliminary achievements and will help to drive a dramatic revolution in clinical practices for the medical community. Through these advances, the diagnosis and treatment of many diseases will achieve many breakthroughs. This project is just beginning and requires a great deal of time and money. Precision medicine also requires extensive collaboration. Ultimately, these difficulties can be overcome. We should realize that precision medicine is good for patients, but there is still a long way to go.

References

[1] https://www.whitehouse.gov/blog/2015/01/21/precision-medicine-improving-health-and-treating-disease

[2] https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

[3] http://finance.china.com.cn/industry/medicine/yyyw/20150326/3024172.shtml

[4] Katsnelson A (2013) Momentum grows to make ‘personalized’ medicine more ‘precise’. Nature Medicine 19: 249.

[5] 11. White, T. Hadoop: The definitive guide. O’Reilly Media, Inc; 2012. 12. Hortonwork. A Modern Data Architecture with Apache Hadoop, The Journey to a Data Lake. Hortonworks; 2014. Tech Rep 13.

[6] Jadhav, AS.; State, W. Online Information Searching for Cardiovascular Diseases: An Analysis of Mayo Clinic Search Query Logs. 2014. 14.

[7] Aronson, R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program; Proceedings/AMIA ... Annual Symposium AMIA Symposium; Jan. 2001; p. 17-21.

[8] https://www.nature.com/news/personalized-medicine-time-for-one-person-trials-1.17411

[9] https://www.bioprocessonline.com/doc/big-data-analytics-the-next-evolution-in-drug-development-0001

[10] Hudson, Kathy, Rick Lifton, and B. Patrick-Lake. "The precision medicine initiative cohort program—Building a Research Foundation for 21st Century Medicine." Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director (2015).

[11] Liu, Xiaoqin, Xin Luo, Chunyang Jiang, and Hui Zhao. "Difficulties and challenges in the development of precision medicine." Clinical genetics (2019).

[12] Le Tourneau, Christophe, Jean-Pierre Delord, Anthony Gonçalves, Céline Gavoille, Coraline Dubot, Nicolas Isambert, Mario Campone et al. "Molecularly targeted therapy based on tumor molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomized, controlled phase 2 trial." The lancet oncology 16, no. 13 (2015): 1324-1334.

[13] Garrido-Castro, Ana C., Nancy U. Lin, and Kornelia Polyak. "Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for

Treatment." Cancer discovery 9, no. 2 (2019): 176-198.

[14] Abramovitz, Mark, Casey Williams, Pradip K. De, Nandini Dey, Scooter Willis, Brandon Young, Eleni Andreopoulou et al. "Precision Medicine Clinical Trials: Successes and Disappointments, Challenges and Opportunities–Lessons Learnt." In Predictive Biomarkers in Oncology, pp. 593-603. Springer, Cham, 2019.

[15] Dimitrov, Dimiter V. "Medical internet of things and big data in healthcare." Healthcare informatics research 22, no. 3 (2016): 156-163.

[16] Soman, A. K. Cloud-based solutions for healthcare IT. CRC Press, 2011.

[17] Lillie, Elizabeth O., Bradley Patay, Joel Diamant, Brian Issell, Eric J. Topol, and Nicholas J. Schork. "The n-of-1 clinical trial: the ultimate strategy for individualizing medicine?." Personalized medicine 8, no. 2 (2011): 161-173.

[18] Leijten, Emmerik F., Timothy R. Radstake, Iain B. McInnes, and Johannes W. Jacobs. "Limits of traditional evidence-based medicine methodologies exemplified by the novel era in psoriatic arthritis drug development." Expert review of clinical immunology just-accepted (2019).

Bibliography

Jafar Ali Ibrahim. S has completed his Bachelor of Technology in Information Technology from Syed Ammal Engineering College, Ramanathapuram, affiliated to Anna University, Chennai, Tamilnadu, India, and Master of Technology degree in Computer Systems and Networks from Dr. MGR Educational and Research Institute, Chennai, Tamil Nadu, India. He possesses nearly 5 years of Industrial experience in the field of IT Security, Bio Metric Deployment, IT Administration, Network Security, Web Services and Security, Service Oriented Architecture environment etc. He has presented the paper in 11 reputed International conferences and published fourteen papers in reputed journals. His research interest includes Cloud Computing, Clinical Research, and Medical Informatics, Internet of Things, Machine Learning, Ontology Development, Big Data, and Data mining. He visited countries like Japan, Malaysia, Singapore, Srilanka, and Gulf countries for his research activities. Right Now he is doing his Ph.D. as a full-time Doctoral Research Fellow in the area of Translational Clinical Informatics at Anna University, Chennai, Tamilnadu, India.

Dr. M. Thangamani possesses nearly 23 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services, and open source software. She has published nearly 80 articles in refereed, indexed, SCI Journals, books, and book chapters and presented over 67 papers in national and international conferences in the above field. She has delivered more than 60 Guest Lectures in reputed engineering colleges and reputed industries on various topics. She has got the best paper awards from various education-related social activities in India and Abroad. She has received the many National and International Awards. She continues to actively serve the academic and research communities and presently guiding nine Ph.D. Scholars under Anna University. She is on the editorial board and reviewing committee of leading research SCI journals. She has on the program committee of top international data mining and soft computing conferences in various countries. She is also Board member in Taylor & Francis Group and seasonal reviewer in IEEE Transaction on Fuzzy System, the international journal of advances in Fuzzy System and Applied mathematics and information journals. She has an organizing chair and keynotes speaker in international conferences in India and countries like California, Dubai, Malaysia