Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients

Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin
Institute of Technology, CWDS, UW Tacoma
{kiyana,mnaren,ankurt, senjutib, scchin}@u.washington.edu

Brian Muckian
Multicare Health System
Tacoma, Washington
brian.Muckian@multicare.org

***Abstract*—Developing holistic predictive modeling solutions for risk prediction is extremely challenging in healthcare informatics. Risk prediction involves integration of clinical factors with socio-demographic factors, health conditions, disease parameters, hospital care quality parameters, and a variety of variables specific to each health care provider making the task increasingly complex. Unsurprisingly, many of such factors need to be extracted independently from different sources, and integrated back to improve the quality of predictive modeling. Such sources are typically voluminous, diverse, and vary significantly over the time. Therefore, distributed and parallel computing tools collectively termed big data have to be developed. In this work, we study big data driven solutions to predict the 30-day risk of readmission for congestive heart failure (CHF) incidents. First, we extract useful factors from National Inpatient Dataset (NIS) and augment it with our patient dataset from Multicare Health System (MHS). Then, we develop scalable data mining models to predict risk of readmission using the integrated dataset. We demonstrate the effectiveness and efficiency of the open-source predictive modeling framework we used, describe the results from various modeling algorithms we tested, and compare the performance against baseline non-distributed, non-parallel, non-integrated small data results previously published to demonstrate comparable accuracy over millions of records.**

Keywords: Healthcare; Knowledge-Discovery; Risk Prediction;

I. INTRODUCTION

Hospital readmission is expensive and generally preventable [1]. Reducing preventable readmission is considered a key quality of care parameter that is deemed measurable. Yet, it is still challenging to develop accurate predictive models to predict such risk and the importance of factors that contribute to readmission due to the diversity of data sources even within a single large hospital. Add to this the aspiration of obtaining a holistic view of cause for readmissions by integrating socio-economic parameters and external data with existing clinical data, and this problem becomes even more challenging and complex requiring significant advances in data integration, discretization, normalization and data organization to name a few.

At the same time, predicting risk of re-hospitalization for chronic and potentially fatal diseases such as congestive heart failure can result in significant cost savings and improvement of care at many hospitals. A key question being asked today in health informatics is how big data healthcare implementations can help correlate and collate insights across various heterogeneous data sources to enable a better understanding of issues such as quality of care particularly for chronic conditions that lead to repetitive readmissions.

It was prohibitively difficult to store, manage and mine large volumes of structured and semi-structured health record datasets prior to the recent advances in big data infrastructure [2]. One of the emergent abilities of new shared nothing, distributed, and parallel computing infrastructure is the ability to perform similar operations on large amounts (petabytes) of data. These infrastructures are evolving to be able to process such large volumes, high velocity, and diverse types of data (variety of data) due to the inherent nature of bringing computation closer to where the data is which is unlike the prior paradigm of having to move data around for large computations to happen. Within the healthcare informatics setting, this ability to process large amount of diverse unstructured, semi-structured, and structured data enables clinical informatics to develop new insights and discovery new knowledge by combining data from various sources. Such sources can be internal as well as external to the electronic health record (EHR) and may include millions of rows and hundreds of attributes, which can be leveraged for predictive modeling.

New tools and programming paradigms for such data intensive applications leverage the distributed computation model. Apache Hadoop [3] is one such distributed framework that implements a computational paradigm MapReduce[4], where the application is divided into many small fragments of work, each of which may be executed or re-executed on a number of compute nodes in a cluster of data intensive distributed applications. In conjunction with a distributed system such as Hadoop, the Apache Mahout [5] framework provides a useful set of machine learning libraries for implementing modeling tasks such as classification and clustering though there is significant need to develop advanced domain specific implementations of these algorithms. Mahout was designed to work in conjunction with Hadoop to scale for large datasets and compute clusters. Furthermore, distributed query processing solutions such as Hive [6] and Cassandra [7] are now available for distributed query processing and exploratory analyses, though very few case studies are available that demonstrate their use in the healthcare setting.

978-1-4799-1293-3/13/$31.00 ©2013 IEEE

Such tools enable important quality of care metrics to be developed across hundreds of dimensions. In this work the focus is on demonstrating how our implementation using current big data tools leveraged this ability to manage millions of events efficiently to develop accurate predictive models for estimating the risk of readmission of congestive heart failure patients.

Congestive Heart Failure (CHF) has been identified as one of the leading causes of hospitalization, especially for adults older than 65 years of age [8]. Furthermore, studies show that CHF is one of the primary reasons behind readmission within a short time-span [9]. Based on the 2005 data of Medicare beneficiaries, it has been estimated that 12.5% of Medicare admissions due to CHF were followed by readmission within 15 days, accounting for about $590 million in health care costs [10]. The Center for Medicare and Medicaid Services (CMS) has started using the 30 day all cause heart failure readmission rate as a publicly reported efficiency metric. All cause 30 day readmission rates for patients with CHF have increased by 11% between 1992 and 2001 [11].

In collaboration with Multicare Health Systems (MHS) [12], we embark on designing big data solutions for predicting risk of readmission for the real world CHF patient records provided by MHS. The necessity of adopting big data solutions in this work is two fold – 1) first, we augment our patient dataset by extracting and integrating additional important factors (such as income) from National Inpatient Dataset (NIS) using Hive and Cassandra. 2) Then, we formalize the problem as a supervised learning task, and design multiple classification algorithms [13] for readmission prediction using distributed analytics platform Mahout[6]. In our experimental analysis, we observe that a 16000 patient dataset with 76 attributes take 32.75 seconds to run giving accuracy of about 79% in Mahout, whereas a 1.5 million patient dataset with the same set of attributes takes 4 minutes 35 seconds (compared to 4 hours 5 minutes, when run in a non-distributed analytics platform) and has 81% accuracy, and finally a 3.2 million patient dataset takes 7 minutes 21 seconds to run (it does not even run in a non-distributed environment) with the same 81% accuracy, demonstrating the effectiveness of our proposed solutions.

Our primary contributions are:

a) We initiate the study of predicting 30-day risk-of readmission problem for CHF patients using big data solutions.
b) We propose distributed solution for information extraction and integration, and propose solutions for predictive modeling using distributed classification models.
c) We conduct a comprehensive set of experiments that demonstrate the quality of the obtained solutions for the risk prediction problem, as well as its scalability aspects on large datasets.

The rest of the paper is organized as follows: In the next section, we describe our proposed big data solutions for information extraction and integration. Then, we discuss our predictive modeling techniques in conjunction with the big data infrastructure. Our comprehensive experimental results are presented in the experiments and results section after that. An overview of current state of the art approaches for risk of readmission and healthcare data integration within this context are covered in the related work section followed by our conclusion and a brief description of some future research directions.

II. Data Extraction and Preprocessing Using Big Data Framework

A. Data Extraction

Real world clinical data is noisy and heterogeneous in nature, severely skewed, and contains hundreds of relevant yet sometimes correlated attributes. This data resides in multiple databases such as individual EMRs, lab and imaging systems, physician notes, medical correspondences, claims, CRM systems, and hospital finance department servers. The collection, integration, and analysis of such big, complex, and noisy data in healthcare are a challenging task. For this reason, healthcare information systems can be considered as a form of big data not only for its sheer volume, but also for its complexity and diversity which makes traditional data warehousing solutions prohibitively cumbersome and ill-suited for large scale data exploration and modeling.

In this section, we study how a big data framework can be leveraged to extract and preprocess data. The focus of the next section will be subsequent predictive modeling. We will leverage Hadoop as our big data framework to archive performance, scalability and fault tolerance for our task at hand. Hadoop is a popular open-source map-reduce implementation, which is being used as an alternative to store and process extremely large data sets on commodity hardware. Hadoop is designed to scale up from single servers to hundreds of compute nodes, each offering local computation and storage capabilities within Hadoop.

However, Hadoop provides no query functionality. In addition, selection methods in Hadoop are comparatively slower than in most DBMS. Thus a processing framework on top of MapReduce solution is also needed to simulate a scalable data warehouse. To achieve this goal, we use Hive [6] as an open-source data warehousing solution built on top of Hadoop. Hive supports queries expressed in a SQL-like declarative language - HiveQL, which are compiled into map-reduce jobs that are executed using Hadoop. In addition, HiveQL enables users to plug in custom map-reduce scripts into queries. Hive has 2 main user interfaces of CLI (command line) and Web UI for access to the data using a SQL like construct. The process is as follows: first the healthcare data such as raw patient event logs, or structured electronic medical records can be stored as flat files on various nodes. These will then become accessible (i.e loaded) into HDFS (Hadoop File System). Then one has to manually invoke Hive commands to create appropriate tables and develop the schema so that data can be structured and appropriately queried. The meta-data of

Hive includes the name, column, partition, and properties of the tables, and it is usually stored in a small relational database table using MySQL. To extract relevant information from the Hive schema, we can run queries using an interface similar to standard SQL which actually gets converted to programmatic constructs and executes as multiple MapReduce jobs.

In addition to Hive, Cassandra is another popular open source big data tool for distributed data management that we leverage in order to make the data extraction process faster. It can handle very large amounts of data spread out across many commodity servers with no single point of failure (due to replication). In comparison with Hive, Cassandra provides a structured key-value store with tunable consistency. Due to its ring architecture, it is massively scalable. Data is replicated to multiple nodes to protect from loss during node failure. Cassandra also offers flexible schema-less data modeling by offering the organization of a traditional RDBMS table layout combined with the flexibility and power of no stringent structure requirements.

B. Data Integration

Many measures of healthcare delivery or quality are not publicly available at the individual patient or hospital level largely due to privacy restrictions, legal issues or reporting norms. Instead, such measures are provided at aggregate level with varying granularity such as state-level, county-level or city-level. For example, average income is typically available by zip-code, whereas death ratio is available by city, or average smoking rate by country, through a variety of publicly available datasets. Although these aggregated statistics cannot reconstruct the underlying individual-level data, these aggregated data can be combined with individual data to produce more informative models. To integrate such data from different sources, in this paper we propose a simple but effective clustering based technique. For example suppose we have two datasets A and B. The dataset A contains income data based on the zip-code, and we want to add this factor to dataset B. To achieve this, the dataset A (including income data based on the zip-code) is divided into a set of clusters using clustering method based on some common features between dataset A and B. Then, the average income is calculated for each cluster. In the next step, each record of B dataset is assigned to a cluster that is most similar to it (based on distance function on common set of features). Finally, income values of the records in B are patched up with the plausible values generated from its respective cluster (Average value).

III. BIG DATA FRAMEWORK FOR RISK OF READMISSION PREDICTIVE MODELING

Predictive models are appropriate for various kinds of clinical risk assessments in health care domain. Clinical risk calculators and risk assessment tools provide information about a person's chance of having a disease or encountering a clinical event [14]. Such tools are useful to educate patients as well as healthcare providers to monitor the development of health conditions. Risk calculators are commonly used for diseases like cancer, diabetes, heart disease, and stroke etc. Developing predictive modeling solutions for such disease related risk of readmissions is extremely challenging in healthcare informatics due to high dimensionality and large volume of the data that is increasingly becoming available within hospital systems. In this paper, the focus is on demonstrating how clinical risk calculator tools can be augmented and scaled using a big data infrastructure implementation.

As mentioned earlier, Risk of Readmission (RoR) prediction for congestive heart failure is challenging [15]. Congestive Heart Failure (CHF) is one of the leading causes of hospitalization, and studies show that many of these admissions are readmissions within a short window of time. Identifying CHF patients who are at a greater risk of hospitalization can guide implementation of appropriate plans to prevent these readmissions. During the initial hospitalization (either during admission or discharge) of a patient, if her risk of readmission (RoR) within a given timeframe (such as, within 30 days or 60 days) could be calculated, it may in turn lead to developing improved post-discharge planning for the patient. Furthermore, such insights may guide health care providers to develop programs to improve the quality of care and administer targeted interventions - thus reducing the readmission rate and the cost incurred in these readmissions. This can also facilitate proper resource utilization within the hospitals.

Formally, the problem is formulated as a supervised learning problem, especially as a binary classification task. The class of a patient is Readmission if the elapsed period between the last discharge and next admission is smaller or equal to 30, and No Readmission else. Developing predictive modeling solutions for ROR prediction is extremely challenging. It involves integration of socio-demographic factors, health conditions, disease parameters, hospital care quality parameters, and a variety of variables specific to health care providers making the task immensely complex. To tackle this complexity, we leverage the power of Mahout as a big data solution for data analytic tasks to archive better scalability in term of time and resources.

Mahout is a machine learning based algorithm library intended to run as Apache MapReduce jobs on the Hadoop cluster in order to be scalable to reasonably large data sets. The advantage of Mahout over other approaches and machine learning tools such as R becomes striking as the number of training examples gets extremely large. This is the scenario, we face in healthcare domain collecting large amount of data 24/7 in an organization's health information and clinical systems. The reason that Mahout has an advantage with larger data sets is that as input data increases; the time or memory requirements for training may not increase linearly in a non-scalable system. In general, the classification algorithms in Mahout require resources that increase no faster than the number of training or test examples, and in most cases the computing

resources required can be parallelized. This allows us to trade off the number of computers used against the time the problem takes to solve.

Fig. 1 shows the framework of using Mahout for RoR prediction. The training data should first load into Hadoop file system (HDFS). In the second step, the raw data should be preprocessed into classifiable data then classifiable data by selecting predictor and target variables and identifying each variable type (numeric, categorical, text and so on) then encode them as vectors, the style of input required by Mahout classifiers. In the third step, the classification algorithm should be selected. The algorithms in Mahout all share scalability. In this paper, we first use random forest because it can work with all types of predictor variables. Moreover, It has high overhead for training thus costly for traditional tools such as R but offers complex and interesting classifications and can handle nonlinear and conditional relationships in data better than other techniques.

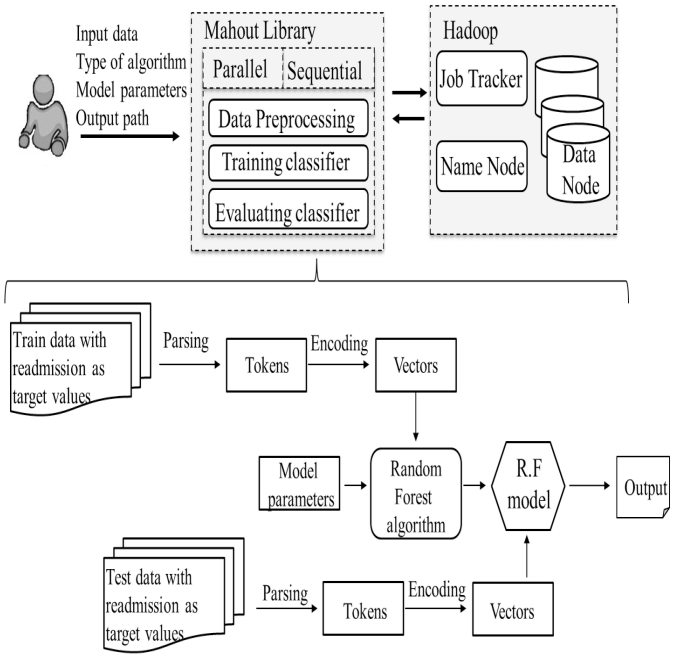

Figure 1. Predictive modeling using Mahout Framework

IV. RESULTS AND DISCUSSION

In this Section we present our comprehensive experimental results. We investigate both quality and the scalability aspects of our proposed solutions for predicting risk of readmission for CHF patients.

A. Datasets Description

Our experiments are mainly conducted using dataset from Multicare Health System (MHS). Hospital encounters of the patients with discharge diagnosis of CHF (either primary or secondary) are identified as the potential index for the CHF related admissions. We primarily consider patients with a discharge diagnosis of ICD9-CM for this purpose, as listed in Table I.

TABLE I. THE ICD-9 CM CODES FOR CHF

ICD-9 CM codes	Description
402.01	Malignant hypertensive heart disease with heart failure
402.11	Benign hypertensive heart disease with heart failure
402.91	Unspecified hypertensive heart disease with heart failure
404.01	Malignant hypertensive heart and kidney disease with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
404.03	Malignant hypertensive heart and kidney disease with heart failure and chronic kidney disease stage V or end stage renal disease
404.11	Benign hypertensive heart and kidney disease with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
404.13	Benign hypertensive heart and kidney disease with heart failure and chronic kidney disease stage V or end stage renal disease
404.91	Unspecified hypertensive heart and kidney disease with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
404.93	Unspecified hypertensive heart and kidney disease with heart failure and chronic kidney disease stage V or end stage renal disease
428.XX	Heart Failure codes

Our entity of observation is each CHF hospital encounter and we consider only the admissions when a patient is discharged to home to exclude inter hospital transfers. Admissions encountering in-hospital deaths are not included in our analysis because we are more interested in predicting readmissions. We calculate the days elapsed between the last discharge due to CHF and next admission in order to identify if the readmission has occurred within 30 days. The dataset consists of CHF hospitalization for patients discharged since 2009. It provides information of 6739 patients diagnosed with CHF and number of hospital encounters generated by these patients during 2009-2013 is 15696

Given the hospital encounter records of every patients, a record has been labeled as \readmission= yes" (or class 1), if this hospitalization is within 30 days of discharge of an earlier index hospitalization due to CHF, or `readmission= no" (or class 0) otherwise. As stated earlier, 30 day is chosen because it is a clinically meaningful time-frame for hospitals and medical communities to take action to reduce the probability of readmission.

Hospital readmission due to CHF is a complex phenomenon governed by multiple factors (i.e, attributes/variables) because of the complexity and uniqueness of the domain. One of our major challenges before classification task is to determine the subset of attributes (i.e., predictor variables) that has significant impact on readmission of patients from the myriad of attributes present in the data set. Based on inputs from domain experts and literature review, Predictor variables can be divided into clinically relevant categories: socio-demographic, vital signs, laboratory tests, discharge

disposition, medical comorbidity and other cost related factors, like length of stay. Many of the vital signs and even laboratory tests are considered both at admission and discharge time.

The Nationwide Inpatient Sample (NIS) is the second dataset used in this experiment to augment the MHS dataset. NIS dataset, which developed as part of the Healthcare Cost and Utilization Project (HCUP is the largest publicly available all-payer inpatient care database in the United States. It contains data from approximately 8 million hospital stays each year; however, it excludes data elements that could directly or indirectly identify individuals so it can not directly integrated to MHS dataset.

The NIS includes more than 100 clinical and nonclinical features for each hospital stay including Primary and secondary diagnoses and procedures, Admission and discharge status Patient demographics (e.g., gender, age, race, median income for ZIP Code) , Expected payment source, Total charges, Length of stay and Hospital characteristics (e.g., ownership, size, teaching status).Considering the large size and high dimensional characteristics of NIS data, it is a prominent example of Big data in healthcare domain.

B. Data Extraction and Integration

As mentioned in previous section, NIS data with 8 million records and more than 100 features is consider as a big data which selecting and extracting data from that is time-consuming. To extract data from NIS, Hive framework is leveraged. The main goal is to integrate income value which is available in the NIS data to MHS data. However due to security reason, it is not possible to map patients in NIS to patients in MHS data. Moreover, the medium income is available based on zip code patient, which is also masked. The only information to find MHS patients in NIS data is the hospital zip code. So in the first step we select the patients in NIS that are hospitalized in MHS facilities using MHS hospitals zip code and Hive framework.it took only 55 seconds for MapReduce job in hive to complete the query and 25 seconds in Cassandra, however, running the same query in traditional RDBMS will take significantly longer. In the next step we used a set of common factors between NIS dataset and MHS that can be correlated to income. We use these factors for the clustering method mentioned in data integration of section II. We choose the age, gender and also elective hospitalization as three variables that are highly correlated to income and can be used for clustering purpose. The table II shows the correlation of these factors to income using chi-square test.

TABLE II. CORRELATION ANALYSIS OF COMMON FACTORS AND INCOME

Factors	**Metrics for Correlation Analysis**	
	X-squared	***p-value***
Age	447.5904	< 2.2e-16
Gender	17.2379	0.001738
Elective Hospitalization	359.0019	< 2.2e-16

Kmeans [13] is used as clustering method to segment selected data from NIS. Then the average income is calculated for each cluster. Now we are able to map each record of data in MHS to closest cluster based on Euclidian distance function.). Finally, income values of the records in MHS dataset are patched up with the plausible values generated from its respective cluster (Average value). Based on the result of correlation analysis, there exist a negatively correlation between the average income and readmission risk. Now average income can be used as a predictor variable for RoR prediction**.** This generic scenario can be used for other attributes in order to augment the original dataset (MHS) with more variables from public and available dataset.

C. Predictive Model Quality Result

Model quality was assessed through common model quality measures such as Accuracy, Precision, Recall and Area Under the Curve (AUC) value. Depending on the final goal of the RoR prediction, the different evaluation measures are less or more appropriate. The precision is important if there is a high cost related to falsely predicting patients to belong to the class Readmission. Recall is relevant if the detection of patients that belong to Readmission is the main goal. The accuracy is the traditional evaluation measure that gives a global insight in the performance of the model. The AUC measure is typically interesting when the problem is imbalanced such the situation we deal with in RoR prediction (it is observed that the labeled dataset is highly skewed - i.e., the number of instances with No Readmission label significantly outnumbers the number of instances with class label Readmission.

Table V shows the result of RoR prediction on original dataset. Logistic regression and random forests are applied to data as two common classification models. We also compare all results with the Yale model as baseline method. In this work, a hierarchical logistic regression model was developed to predict 30-day readmission risk for patients hospitalized with heart failure using some demographic and comorbidity variables. As our data at hand is different from the one considered in the Yale Model, our comparison primarily relies on the basis of the attributes suggested by the Yale model. To circumvent the imbalanced problem, we leverage over sampling method. This technique alters the class distribution of the training data so that both the classes are well represented. Oversampling works by resampling the rare class records so that the resulting training set has an equal number of records for each. The positive effect of over-sampling can be observed in improving the recall metric for logistic regression model.

D. Predictive Model Result for Scalability

The objective of the experiments in this section is to show to what extent big data solution can lead to time efficiency and scalability. This paper describes a benchmark study of various scenarios that are created by applying random forest as classification algorithm to compare the performance of two-open source software packages: R as traditional statistical tool and Apache Mahout as a big data solution for machine learning. Each scenario was evaluated for model quality and time efficiency. Table III shows the general setup for each software platform.

TABLE III. GENERAL SETUP FOR SOFTWARE PLATFORM

Software	System Details	
	Appliance	*Available RAM*
R 3.0.0	Linux server	16GB
Mahout 0.7	Two-Node Hadoop Cluster	2GB(each node)
	Three-Node Hadoop Cluster	
	Four-Node Hadoop Cluster	

Our original data in MHS contains only 15696 records for CHF patients admission in Multicare hospitals which cannot be considered as a big data, however, since our aim is to have a general solution that can be applicable to any data size for RoR prediction, we scaled up the original data linearly several times to show how big data framework outperforms in comparison with traditional systems when the training set becomes larger. The scaled data has created an increased demand for memory and processing task needed to predict RoR. Table IV shows the five different scenarios of data size.

TABLE IV. SCENARIOS FOR DATASET

Number of Tuples in Dataset				
Scenario 1	*Scenario 2*	*Scenario 3*	*Scenario 4*	*Scenario 5*
15,696	114,862	900,000	1,665,866	3,271,716

In order to directly compare the modeling results, parameters were chosen consistently across software. For example, in random forest model, the number of trees equals to 120 and each tree has 15 variables at random. The data were split into a 70/30 training and test partition, for all the scenarios. In these set of experiments, we ignore the logistic regression result because the execution mode for logistic regression is sequential so it cannot be used to show the parallelization benefit of big data solution. Table VI shows the result of random forest model in both R and Mahout Framework for all the scenarios. Random forest algorithm trains an enormous number of simple classifiers and uses a voting scheme to get a single result. The Mahout parallel implementation trains many classifiers in the model in parallel. We choose Random forest since this approach has somewhat unusual scaling properties. Because each small classifier is trained on some of the features of all of the training examples, the memory required on each node in the cluster will scale roughly in proportion to the square root. Thus the size of vectors is large and larger vectors consume more memory and slow down the training. That's the main reason that R was not able to run random forest in scenario 5 because the data is too large to fit into main memory. These set of experiment show the main advantage of Mahout, which is its robust handling of extremely large and growing data sets.

The Fig. 2 shows the training time of random forest model in R and Mahout Platforms.it can be observed that when the number of training examples is relatively small, traditional data mining tools work even better than Mahout (scenario 1). However, as the training size increases, Mahout's scalable and parallel algorithm is significantly better with regard to time in comparison with traditional statistical tools such as R. The increased time required by non-scalable algorithms is often due to the fact that they require unbounded amounts of memory as the numbers of training examples grow.

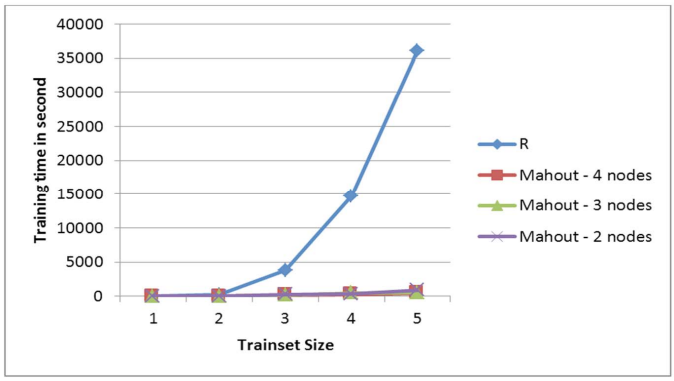

Figure 2. Train time for random forest model

Fig. 3 compares the performance of Mahout Framework for different number of nodes for Hadoop cluster. The parallelization power of mahout can be easily observed from this diagram. When the number of training examples is relatively small, Two-machine Hadoop cluster works even better than three and four-nodes but for larger training data, the training time significantly decreases when we increase the number of nodes in Hadoop cluster.

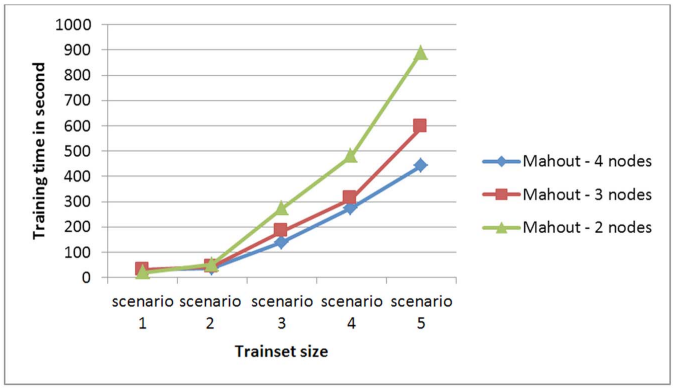

Figure 3. Training time for Mahout with different number of nodes

V. BACKGROUND AND SIGNIFICANCE

Preventing hospitalization is a prominent factor to reduce patient morbidity, improve patient outcomes, and curb health care costs. An increasing body of literature attempts to develop predictive models for hospital readmission risks [1,8,12,14,15,16 ,17,18,19,20,21,22,23,24]. These studies range from all-cause readmissions to readmission for specific diseases such as heart failure, pneumonia, stroke, and asthma. Each of these models exploits various predictor variables assessed at various times related to index hospitalization (admission, discharge, first follow-up visit, etc. In another research study [9], a real-time predictive model was developed to identify CHF patients at high risk for readmission within the 30-day timeframe. In this model, some clinical and social factors available within hours of hospital presentation are used in order to have a real-time predictive model. Although the model demonstrated good discrimination for 30-day readmission (AUC 0.72), the dataset size is very small (1372 HF patients). One of the recent studies for predicting 30-day readmission risk for heart failure hospitalization is done in [15]. In this work, administrative claim data is used to build a regression model on 24,163 patients from 307 hospitals. In a recent research study, we have proposed a risk calculator tool [14] that is capable of calculating 30-day readmission risk for Congestive Heart Failure based on incomplete patient data. In a separate research effort [25], we also demonstrate the effectiveness of data preprocessing in 30-day readmission risk prediction problem. Novel predictive modeling techniques for 30-day risk-of-readmission prediction problem are investigated by the authors in [25].

A recent study investigates the impact of big data on healthcare solutions [26]. This study suggests that leveraging the collection of patient and practitioner data could be an important way to improve quality and efficiency of health care delivery. In fact, while the complexity of the domain, due to very high velocity volume and variety of medical data is acknowledged [26], however, the necessity of enabling big data solutions to these problems is mostly overlooked in the previous works. To the best of our knowledge, we are the first one to propose big data solutions for information extraction, information integration, and predictive modeling for 30-day readmission risk prediction problems.

Fortunately there exist tools and programming paradigms for such data intensive applications leveraging distributed computation model. Apache Hadoop is one such distributed framework that implements a computational paradigm MapReduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster data intensive distributed applications, such as Apache Hadoop. In conjunction with a distributed system such as Hadoop, Apache Mahout provides a useful set of machine learning tools that allow data to be classified, clustered, and filtered, and Mahout was designed to work with Hadoop so it is easily scaled up to large dataset and networks. MADlib and Bismark[27,28] are two other useful analytics tools that are designed to analyze structured and unstructured data in parallel. These are excellent tools to enable scalable, sophisticated in-database analytics and have been well adopted by the database-engine developers, data scientists, IT architects and academics.

VI. CONCLUSION

In this work, we study the big data solution for predicting the 30-day risk of readmission for the CHF patients. Our proposed solution leverages big data infrastructure for both information extraction and predictive modeling. We study the effectiveness of our proposed solution with a comprehensive set of experiment, considering quality and scalability. As ongoing work, we aim at leveraging big data infrastructure for our designed risk calculation tool, for designing more sophisticated predictive modeling and feature extraction techniques, and extending our proposed solutions to predict other clinical risks.

VII. ACKNOWLEDGMENT

This work is supported by MHS (grant no: A73191). Additionally, we are thankful to the data architects and the clinicians at MHS for their valuable time and insightful discussions during the initial stage of the study.

TABLE V. RESULT OF RoR PREDICTION ON ORIGINAL DATASET

Models	**Supervised Learning Algorithms**	**Class Imbalance Solution**	**Results for Prediction**				
			Accuracy	***Precision***	***Recall***	***F-measure***	***AUC***
MHS Model	Logistic Regression	-	77.88%	32%	0.69%	1.36%	63.78%
		OS	58.39%	28.90%	61.41%	39.30%	63.24%
	Random Forest	-	77.90%	40.47%	1.48%	3.01%	61.04%
		OS	77.96%	44.44%	1.7%	3.3%	62.25%
Yale Model (Baseline)	Logistic Regression	-	78.03%	33%	0.08%	0.17%	59.72%

TABLE VI. RANDOM FOREST PREDICTION RESULT ON R AND MAHOUT

Dataset	**Platform**	**Results of Prediction**				
		Accuracy	***Precision***	***Recall***	***F-measure***	***Runtime***
Scenario 1	R	77.90%	40.47%	1.48%	3.01%	18.96 sec
	Mahout	78.84%	93.61%	3.83%	7.35%	32.75 sec

Scenario 2	R	82.35%	99.59%	17.78%	30.17%	4.01 min
	Mahout	78.55%	94.51%	1.8%	3.7%	36.65 sec
Scenario 3	R	86.34%	99.87%	37.49%	54.52%	1h 17m
	Mahout	80.91%	91.62%	14.49%	25.02%	2m 20 sec
Scenario 4	R	87.12%	99.88%	40.60%	57.73%	4h 5m
	Mahout	80.98%	88.83%	15.94%	27.02%	4m 35sec
Scenario 5	R	Cannot allocate vector of size 3.9 Gb after 10 hour running				
	Mahout	80.79%	91.48%	13.99%	24.27%	7m21sec

REFERENCES

[1] Donzé J. Aujesky D., Williams D., Schnipper J.L, MD. Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. JAMA Internal Medicine, 173(8):632-638, Apr. 2013.

[2] Manyika J., Chui M., Brown B., and Bughin J. and Dobbs R. Big data: The next frontier for innovation competition and productivity, McKinsey Global Institute, 2012.

[3] The Apache Software Foundation., http://hadoop.apache.org/common/credits.html.

[4] Ghemawat D.J .MapReduce: simplified data processing on large clusters. In: Proc of OSDI, 2004.

[5] Owen S. and Anil R. Mahout in Action. Manning Publications Co., Greenwich, Connecticut, 2010.

[6] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., And Murthy, R .Hive—a warehousing solution over a Map-Reduce framework. In VLDB, 2009.

[7] Wikipedia, http://en.wikipedia.org/wiki/Apache_Cassandra.

[8] Adams K. F., Fonarow G. C., Emerman C. L., LeJemtel T. H., Costanzo M. R., Abraham W. T., Berkowitz R. L., Galvao M., and Horton D. P. Characteristics and outcomes of patients hospitalized for heart failure in the United States: Rationale, design, and preliminary observations from the first100, 000 cases in the acute decompensated heart failure national registry (ADHERE). American Heart Journal, 149(2):209-216, Feb. 2005:

[9] Ross JS, Chen J, Lin Z, Bueno H, Curtis JP, Keenan PS, Normand SL, Schreiner G, Spertus JA, Vidán MT, Wang Y, Wang Y, Krumholz HM. Recent national trends in readmission rates after heart failure hospitalization. Circ Heart Fail, 3:97-103, 2010.

[10] Krumholz H. M., Normand S. L. T., Keenan P. S., Lin Z. Q., Drye E. E., Bhat K. R., Wang Y. F., Ross J. S., Schuur J. D., and Stauer B. D.. Hospital 30-day heart failure readmission measure methodology. Report prepared for the Centers for Medicare & Medicaid Services.

[11] Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, Reed WG, Swanson TS, Ma Y, Halm EA. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Journal of Medical Care, 10:981-988, Feb. 2010.

[12] MULTICARE HEALTH SYSTEM, http://www.multicare.org/.

[13] Han J. and Kamber M.. Data mining: concepts and techniques. Morgan Kaufmann, 2006.

[14] Zolfaghar K., Agarwal J., Sistla D., Chin S., Roy S. B., Verbiest N., Teredesai A., Hazel D., Amoroso P., and Reed L. Risk-o-meter: An intelligent clinical risk calculator. In Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2013.

[15] Hammill BG, Curtis LH, Fonarow GC, Heidenreich PA, Yancy CW, Peterson ED, Hernandez AF. Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. Cardiovasc Qual Outcomes, 4(4):60-67, 2011.

[16] Coleman EA, Parry C, Chalmers S, Min SJ. The care transitions intervention: Results of a randomized controlled trial. Archives ofInternal Medicine, 166(17):1822-1828, Sept. 2006.

[17] Franchi C., Nobili A., Mari D., Tettamanti M., Djade C. D., Pasina L., Salerno F., Corrao S. , Marengoni A., Iorio A., Marcucci M., and Mannucci P. M. Risk factors for hospital readmission of elderly patients.European Journal of Internal Medicine, 24(1):45-51, Jan. 2013.

[18] Harrison P., Hara P., Pope J., Young M., and Rula E.. The impact of post discharge telephonic follows up on hospital readmissions. Popul Health Manag, 14:27-32, 2011.

[19] Hunter T., Nelson J., and Birmingham J.. Preventing readmissions through comprehensive discharge planning. Prof Case Manag., 18:56-63, 2013.

[20] Kaur H. and Wasan S. K.. Empirical study on applications of data mining techniques in healthcare. Journal of Computer Science, 2(2):194-200, 2006.

[21] Koelling T. M., Johnson M. L., Cody R. J., and Aaronson. K. D. Discharge education improves clinical outcomes in patients with chronic heart failure. Circulation, 111(2):179-185, Jan. 2005.

[22] Koh H. C. and Tan. G. Data mining applications in healthcare. Journal of Healthcare Information Management Vol, 19(2):65, 2011.

[23] Krumholz H. M., Amatruda J., Smith G. L., Mattera J. A., Roumanis S. A., Radford M. J., Crombie, P. and Vaccarino V.. Randomized trial of an education and support intervention to prevent readmission of patients with heart failure. Journal of the American College of Cardiology, 39(1):83-89, Jan. 2002.

[24] Ottenbacher K., Smith P., Illig S., Linn R., Fiedler R., and Granger C.. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. Journal of clinical epidemiology, 54(11):1159-1165, 2001.

[25] Meadam N., Verbiest N., Zolfaghar K., Agarwal J., Chin S., Basu Roy S., Teredesai A., Hazel D., Reed L., Amoroso P. Exploring Preprocessing Techniques for Prediction of Risk of Readmission for Congestive Heart Failure Patients. In Data Mining and Healthcare Workshop, in conjunction with the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2013.

[26] Murdoch T., Detsky A., The Inevitable Application of Big Data to Health Care, JAMA. 2013; 1351- 1352. doi:10.1001/jama.2013.393.

[27] Hellerstein, J., Schoppmann F., Wang D. Z., Fratkin E., Gorajek A., Welton C., Feng X., and Kumar A. The madlib analytics library or mad skills. PVLDB 2012.

[28] Feng, X., A. Kumar, B. Recht, and C. R´e, 2012: Towards a unified architecture for in-rdbms analytics. In SIGMOD Conference, pp. 325–336