Big Data Analytics (BDA) for Medical Imaging

A Seminar Report by

UTPAL KANT

Under the guidance of

DR. VINOD KUMAR
&
DR. P. SUMATHI

Department of Electrical Engineering
Indian Institute of Technology Roorkee

TABLE OF CONTENTS

Abstract

Health data volume is expected to grow dramatically in the years ahead. In addition, healthcare reimbursement models are changing; meaningful use and pay for performance are emerging as critical new factors in today's healthcare environment. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits. What exactly is big data? A report delivered to the U.S. Congress in August 2012 defines big data as "large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information". Big data encompasses such characteristics as variety, velocity and, with respect specifically to healthcare, veracity. Existing analytical techniques can be applied to the vast amount of existing (but currently unanalysed) patient-related health and medical data to reach a deeper understanding of outcomes, which then can be applied at the point of care. Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient.

Systems for storing and transmitting digital data are increasing rapidly in size and bandwidth capacity. Data collection projects such as Chronic Obstructive Pulmonary Disease COPDGene (19,000 lung CT scans of 10,000 subjects) offer unprecedented opportunities to learn from large medical image sets, for example to discover subtle aspects of anatomy or pathology only observable in subsets of the population. For this, image processing algorithms must scale with the quantities of available data.

With big data we mean any enormous and multifaceted collection of data (texts, numbers, documents, images, videos etc.) that cannot be analysed by ordinary computing devices and algorithms. Big data, due to their sheer volume and inherent variety, are extremely challenging to manage and hence difficult to understand.

More and more users, and more and more machines generate more unstructured data on a daily basis. Capturing the essence of such massive flow of data is beyond traditional computation facilities and their classical methodologies. Large data centres and intelligent algorithms, embedded within capable and flexible distributed computing environments such as Hadoop, are necessary to make sense of big data. One of the major fields "suffering" from big data, which has been widely neglected so far, is medical imaging. More than approximately two trillion medical images are captured worldwide each year. A large number of these images have to be stored for several years. There is a huge amount of information contained in these images and their annotations (notes on diagnosis, biopsy, treatment etc.). Presently this colossal pool of human knowledge is going untapped. Employing machine-learning algorithms may help us to overcome this barrier and to create the frontier for the 21st century medical imaging.

In this seminar report I will attempt to investigate the following questions: What is big data? What is big data analytics? What is Hadoop and what is its relationship to big Data? Why do medical images constitute big data? What is the role of big data analytics in medical imaging?

Introduction to Big Data

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on." Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connect omics, complex physics simulations, biology and environmental research.

Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte's (2.5×1018) of data are created. One question for large enterprises is determining who should own big data initiatives that affect the entire organization.

Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disk it may fill an entire football field. The same amount was created in every two days in 2011, and in every ten minutes in 2013. This rate is still growing enormously. Though all this information produced is meaningful and can be useful when processed, it is being neglected. 90% of the world's data was generated in the last few years.

What is Big Data?

Big data means really a big data it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks.

What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

- **Black Box Data**: It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.

- **Social Media Data**: Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.

- **Stock Exchange Data**: The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.

- **Power Grid Data**: The power grid data holds information consumed by a particular node with respect to a base station.

- **Transport Data**: Transport data includes model, capacity, distance and availability of a vehicle.

- **Search Engine Data**: Search engines retrieve lots of data from different databases.

- **Health Care Data**: Health care Data Includes Clinical Data, Medical Images Biological and Behavioural Data.

Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

- **Structured data**: Relational data.

- **Semi Structured data**: XML data.

- **Unstructured data**: Word, PDF, Text, Media Logs.

Benefits of Big Data

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us:

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.

- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.

- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

Big Data Technologies

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business.

To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in real-time and can protect data privacy and security.

There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology

Operational Big Data

This include systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.

NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.

Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

Analytical Big Data

This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data.

MapReduce provides a new method of analysing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines.

Big Data Challenges

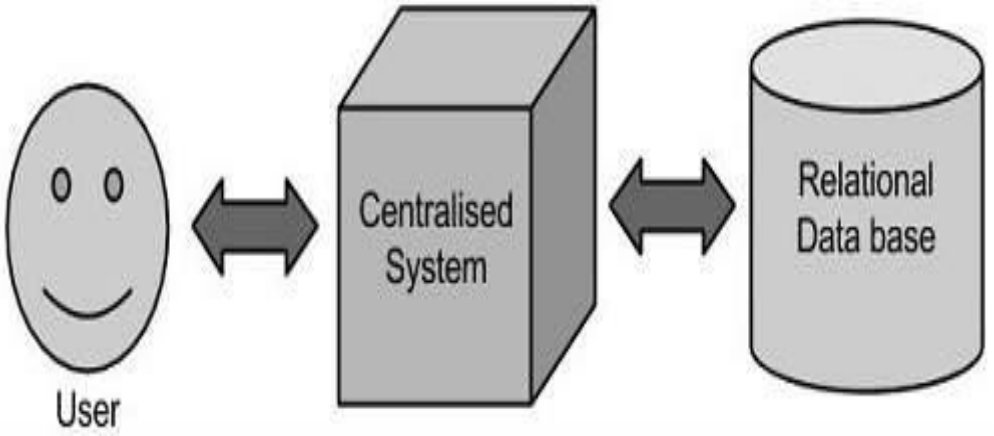
The major challenges associated with big data are as follows:

- Capturing data
- Curation
- Storage
- Searching
- Sharing
- Transfer
- Analysis
- Presentation

To fulfil the above challenges, organizations normally take the help of enterprise servers.

Traditional Approach

In this approach, an enterprise will have a computer to store and process big data. Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated software can be written to interact with the database, process the required data and present it to the users for analysis purpose.



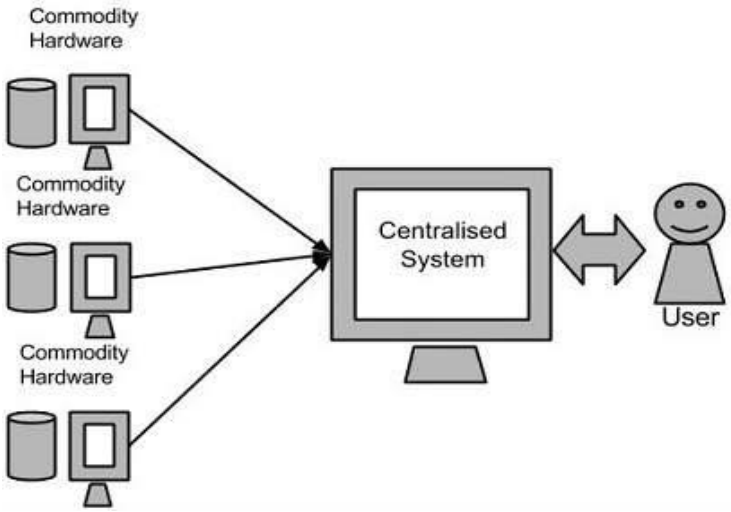
Traditional Approach
Figure 1 [9]

Limitation

This approach works well where we have less volume of data that can be accommodated by standard database servers, or up to the limit of the processor which is processing the data. But when it comes to dealing with huge amounts of data, it is really a tedious task to process such data through a traditional database server.

Google's Solution

Google solved this problem using an algorithm called MapReduce. This algorithm divides the task into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result dataset.

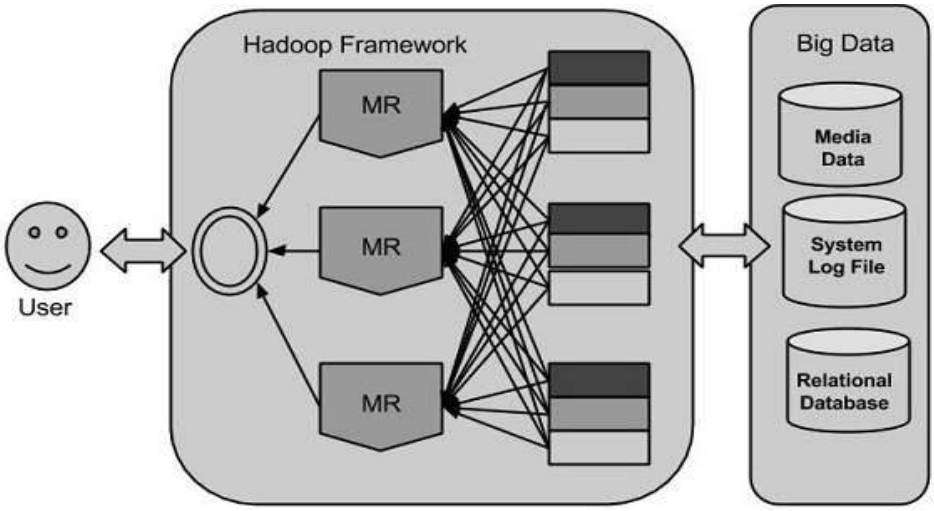


Google File System
Figure 2 [9]

Above diagram shows various commodity hardwires which could be single CPU machines or servers with higher capacity.

Hadoop

Doug Cutting, Mike Cafarella and team took the solution provided by Google and started an Open Source Project called HADOOP in 2005 and Doug named it after his son's toy elephant. Now Apache Hadoop is a registered trademark of the Apache Software Foundation.



MapReduce Architecture
Figure 3 [9]

Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data.

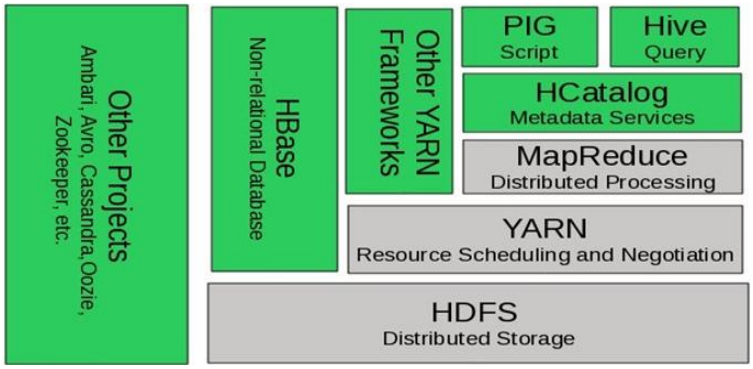
Hadoop Framework

The two classes of Big Data technology are complementary and deployed together, as Hadoop. Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop Architecture

Hadoop framework includes following four modules:

- **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce:** This is YARN-based system for parallel processing of large data sets.
- PIG: Originally developed at Yahoo, 2006, is a High level programming on top of Hadoop MapReduce for scripting in Data analysis problems as data flows.
- Hive: Hive is a Data warehouse software facilitates querying and managing large datasets residing in distributed storage.
- HBase: HDFS data base.
- Spark: Apache Spark™ is a fast and general engine for large-scale data processing its Multi-stage in-memory primitives provides performance up to 100 times faster for certain applications. Spark Allows user programs to load data into a cluster's memory and query it repeatedly, Well-suited to machine learning.



Hadoop Framework
Figure 4 [9]

Since 2012, the term "Hadoop" often refers not just to the base modules mentioned above but also to the collection of additional software packages that can be installed on top of or alongside Hadoop, such as Apache Pig, Apache Hive, Apache HBase, Apache Spark etc.

Hadoop is supported by GNU/Linux platform and its flavours. Therefore, we have to install a Linux operating system for setting up Hadoop environment. In case you have an OS other than Linux, you can install a Virtualbox software in it and have Linux inside the Virtualbox.

Advantages of Hadoop

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware.

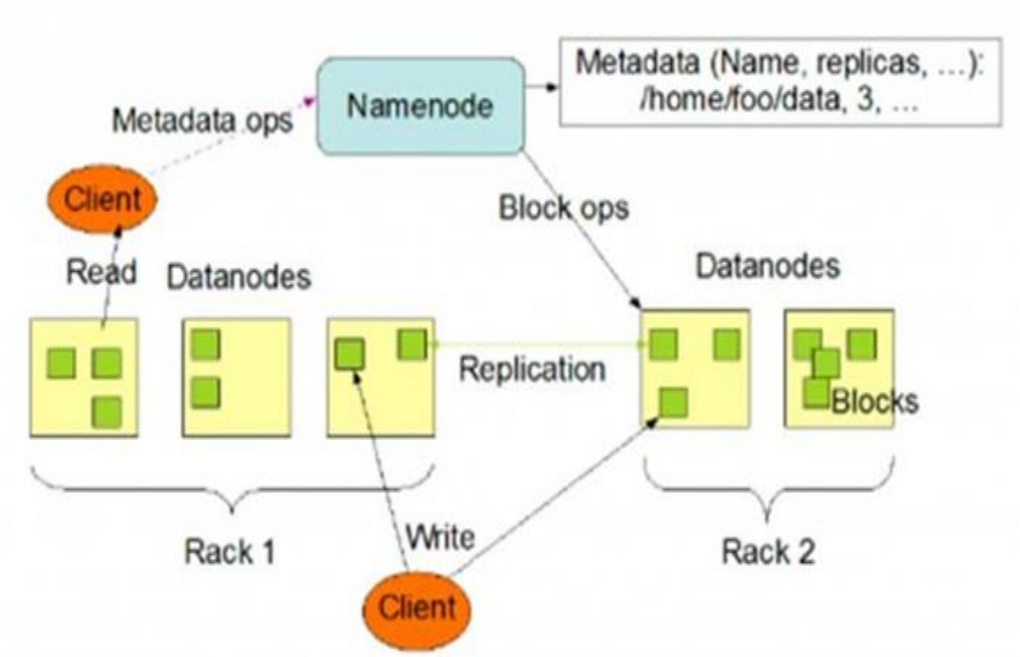
HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing [12].

Features of HDFS

- It is suitable for the distributed storage and processing.
- Hadoop provides a command interface to interact with HDFS.
- The built-in servers of namenode and datanode help users to easily check the status of cluster.
- Streaming access to file system data.
- HDFS provides file permissions and authentication.
- HDFS Architecture

HDFS Architecture

Given below is the architecture of a Hadoop File System.



HDFS Architecture
Figure 5 [8]

HDFS follows the master-slave architecture and it has the following elements.

Namenode

The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software. It is a software that can be run on commodity hardware. The system having the namenode acts as the master server and it does the following tasks:

- Manages the file system namespace.
- Regulates client's access to files.
- It also executes file system operations such as renaming, closing, and opening files and directories.

Datanode

The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. For every node (Commodity hardware/System) in a cluster, there will be a datanode. These nodes manage the data storage of their system.

- Datanodes perform read-write operations on the file systems, as per client request.
- They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.

Block

Generally, the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block. The default block size is 64MB, but it can be increased as per the need to change in HDFS configuration.

Goals of HDFS

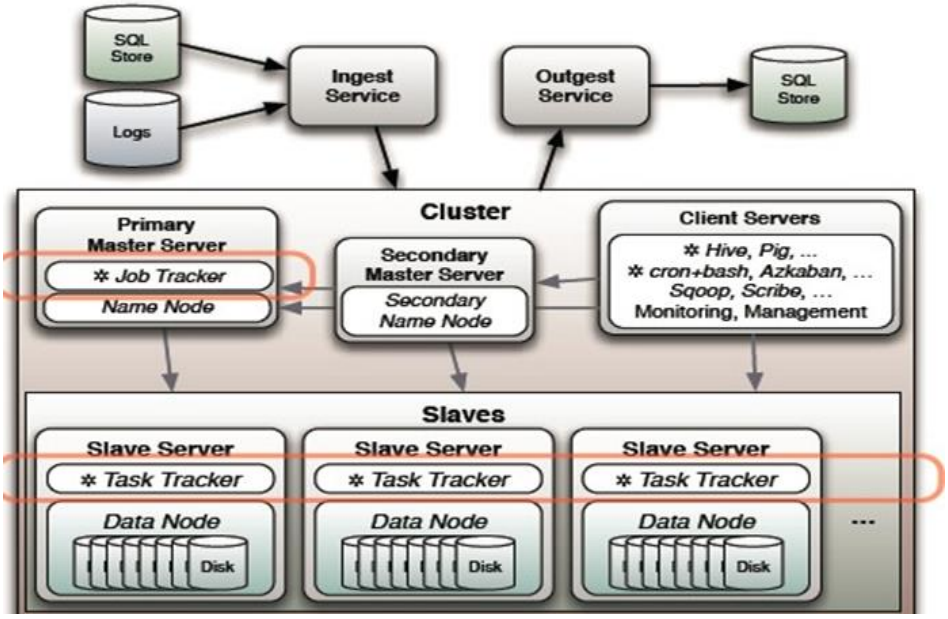
- Fault detection and recovery: Since HDFS includes a large number of commodity hardware, failure of components is frequent. Therefore, HDFS should have mechanisms for quick and automatic fault detection and recovery.
- Huge datasets: HDFS should have hundreds of nodes per cluster to manage the applications having huge datasets.
- Hardware at data: A requested task can be done efficiently, when the computation takes place near the data. Especially where huge datasets are involved, it reduces the network traffic and increases the throughput.
- Streaming Data Access: Applications that run on HDFS need streaming access to their data sets. They are not general purpose applications that typically run on general purpose file systems. HDFS is designed more for batch processing rather than interactive use by users. The emphasis is on high throughput of data access rather than low latency of data access. POSIX imposes many hard requirements that are not needed for applications that are targeted for HDFS. POSIX semantics in a few key areas has been traded to increase data throughput rates.

MapReduce Engine

Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

What is MapReduce?

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job [1].



MapReduce Engine

Figure 6 [8]

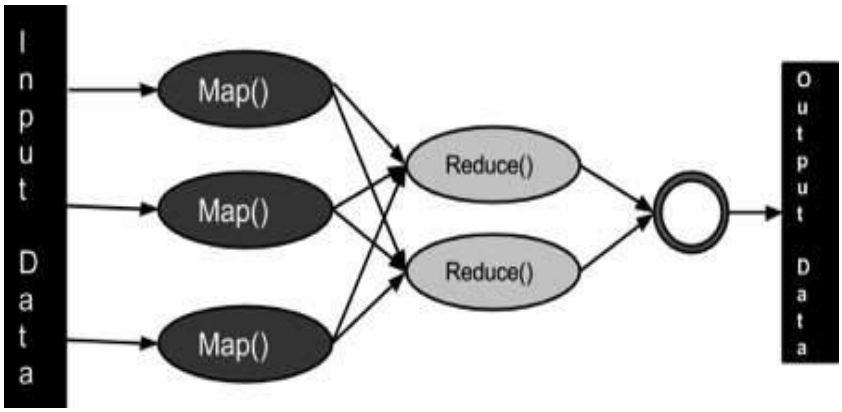
The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

Terminology

- PayLoad - Applications implement the Map and the Reduce functions, and form the core of the job.
- Mapper - Mapper maps the input key/value pairs to a set of intermediate key/value pair.
- NamedNode - Node that manages the Hadoop Distributed File System (HDFS).

- DataNode - Node where data is presented in advance before any processing takes place.
- MasterNode - Node where JobTracker runs and which accepts job requests from clients.
- SlaveNode - Node where Map and Reduce program runs.
- JobTracker - Schedules jobs and tracks the assign jobs to Task tracker.
- Task Tracker - Tracks the task and reports status to JobTracker.
- Job - A program is an execution of a Mapper and Reducer across a dataset.
- Task - An execution of a Mapper or a Reducer on a slice of data.
- Task Attempt - A particular instance of an attempt to execute a task on a SlaveNode.

The Algorithm



MapReduce Algorithm

Figure 7 [9]

- Generally, MapReduce paradigm is based on sending the computer to where the data resides!
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

  Map stage: The map or mapper's job is to process the input data. Generally, the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

  Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.
- During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
- The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
- Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
- After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

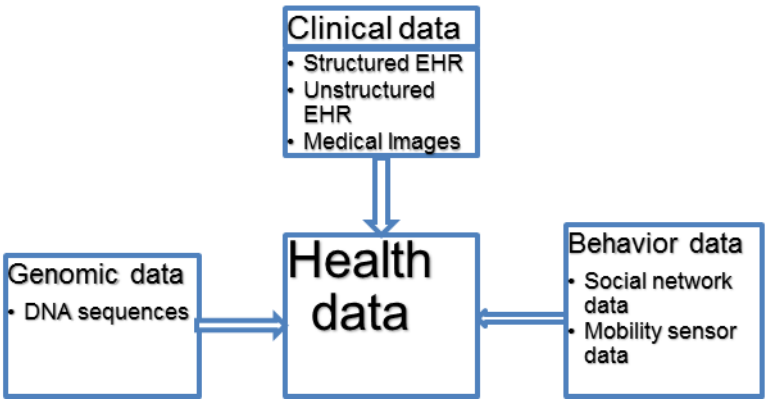
Big Data in Health Care

Health data volume is expected to grow dramatically in the years ahead. In addition, healthcare reimbursement models are changing; meaningful use and pay for performance are emerging as critical new factors in today's healthcare environment. Although profit is not and should not be a primary motivator, it is vitally important

for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits.

What exactly is big data? A report delivered to the U.S. Congress in August 2012 defines big data as "large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information". Big data encompasses such characteristics as variety, velocity and, with respect specifically to healthcare, veracity. Existing analytical techniques can be applied to the vast amount of existing (but cur- rentlyunanalyzed) patient-related health and medical data to reach a deeper understanding of outcomes, which then can be applied at the point of care. Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient.

Figure below shows different form of healthcare data:



Health Care Data

Figure 8

Considered a big data problem in Health Care, in the light of the "5V" definition

Volume: This is probably the big data criterion that current Personalised Healthcare (PH) research fits least well. Although the community wishes to exploit the vast entirety of clinical data records, often there is simply not the level of detail, or depth, that supports the association of parameters in the mechanistic models with the data in the clinical record. The datasets that support these analyses are often very expensive to acquire, and currently the penetration is limited. Never-the less, this is an important area of research, in which the PH community could learn from, and exploit, existing technology from the big data community.

Variety: The variety is very high. In the example at hand we would have clinical data, data from medical imaging, data from wearable sensors, lab exams, and simulation results. This would include both structured and unstructured data, with 3-D imaging posing specific problems of data treatment such as automated voxel classification.

Velocity: Osteoporosis is a chronic condition; as such all patients are expected to undergo a full specialist control every two years, where the totality of the examinations is repeated. Regarding growth rate: If to this we add that the ageing of the population is constantly increasing the number of patients affected, we face growth rates in the order of 55–60% every year.

Veracity: Here there is a big divide between clinical re-search and clinical practice. While data collected as part of clinical studies are in general of good quality, clinical practice tends to generate low quality data. This is due in part to the extreme pressure medical professionals face, but also to a lack of "data value" culture; most medical professionals see the logging of data a bureaucratic need and a waste of time that distracts them from the care of their patients.

Value: The potential value associated with these data is very high. The cost of osteoporosis, including pharmacological intervention in the EU in 2010 was estimated at €37 billion. Moreover, in general, healthcare expenditure in most developed countries is astronomical: the 2013/2014 budget for NHS England was £95.6 billion, with an increase over the previous year of 2.6%, at a time when all public services in the UK are facing hard cuts. In OECD countries, we spend on average USD$3395 per year per inhabitant in healthcare (source: OECD 2011).

The concept of `big data' is not new, however the way it is de ned is constantly changing. Various attempts at defining big data essentially characterize it as a collection of data elements whose size, speed, type and/or complexity require one to seek, adopt and invent new hardware and software mechanisms in order to successfully store, analyze and visualize the data. Healthcare is a prime example of how the three V's of data, velocity (speed of generation of data), variety and volume, are an innate aspect of the data it produces. This data is spread among multiple healthcare systems, health insurers, researchers, government entities, etc. Furthermore, each of these data repositories is inherently incapable of providing a platform for global data transparency. To add to the three V's, the veracity of healthcare data is also critical for its meaningful use towards developing translational research.

Despite the inherent complexities of healthcare data, there is potential and bene t in developing and implementing big data solutions within this realm. A report by McKinsey Global Institute suggests that if US healthcare were to use big data creatively and effectively; the sector could create more than $300 billion in value every year. Two thirds of the value would be in the form of reducing US healthcare expenditure. Historical approaches to medical research have generally focused on the investigation of disease states based on the changes in physiology in the form of a con ned view of certain singular modality of data. Although this approach to understanding diseases is essential, research at this level mutes the variation and interconnectedness that de ne the true underlying medical mechanisms. After decades of technological laggard, the field of medicine has begun to acclimatize to today's digital data age. New technologies make it possible to capture vast amounts of information about each individual patient over a large timescale. However, despite the advent of medical electronics, the data captured and gathered from these patients has remained vastly underutilized and thus wasted.

Important physiological and pathophysiological phenomena concurrently manifest as changes across multiple clinical streams. This results from strong coupling among different systems within the body (e.g., interactions between heart rate, respiration, blood pressure, etc.) thereby producing potential markers for clinical assessment. Thus, understanding and predicting diseases require an aggregated approach where structured and unstructured data

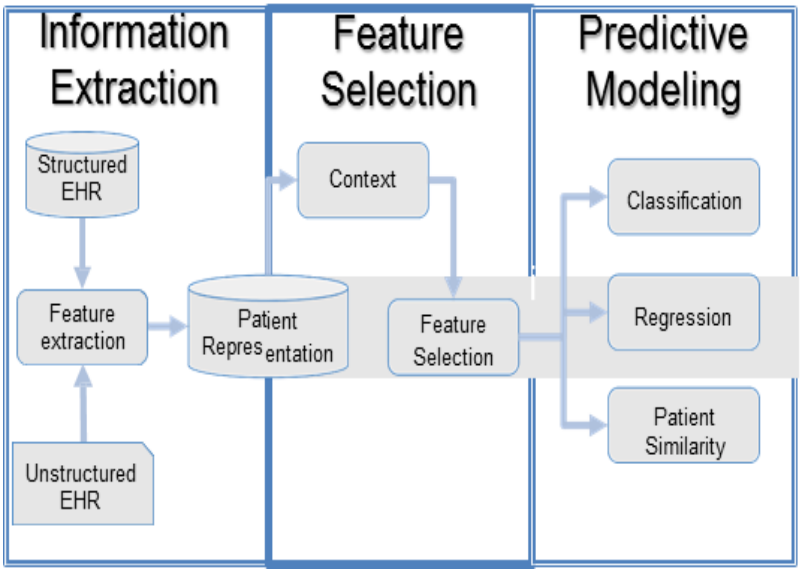
stemming from a myriad of clinical and nonclinical modalities are utilized for a more comprehensive perspective of the disease states. An aspect of healthcare research that has recently gained traction is in addressing some of the growing pains in introducing concepts of big data analytics to medicine. Researchers are studying the complex nature of healthcare data both in terms of characteristics of the data itself as well as in the taxonomy of analytics that can be meaningfully performed on them.

Big Data Analytics in medicine

Big Data Analytics (BDA) in Healthcare can be used as Text mining, Information Extraction

Name Entity Recognition, Information Retrieval from clinical data as well as classification of data such as Clinical text vs. Biomedical text, Biomedical text: medical literatures (well-written medical text) Clinical text is written by clinicians in the clinical settings.

Figure Shows Standard Architecture of Analytic Platform for Healthcare Data:



Analytic Platform

Figure 9

Three main areas of Big Data Analytics in medicine:

Image processing: Medical images are an important source of data frequently used for diagnosis, therapy assessment and planning. Computed tomography (CT), magnetic resonance imaging (MRI), X ray, molecular imaging, ultrasound, photoacoustic imaging, positron emission tomography, computed tomography (PET, CT) and mammography are some of the examples of imaging techniques that are well established within clinical settings. Medical image data can range anywhere from a few megabytes for a single study (e.g. a histology image) to hundreds of megabytes per study (e.g. thin slice CT studies comprising up to 2500+ scans per study). Such data requires large storage capacities if stored long term. It also demands fast and accurate algorithms if any decision assist automation were to be performed using the data. In addition, if other sources of data acquired for each patient are also utilized during the diagnoses, prognosis and treatment processes, then the problem of providing cohesive storage and developing efficient methods capable of encapsulating the broad range of data becomes a challenge.

Signal processing: Similar to medical images, medical signals also pose volume and velocity obstacles especially during continuous, high resolution acquisition and storage from a multitude of monitors connected to each patient. However, in addition to the data size issues, physiological signals also pose complexity of a spatiotemporal nature. Analysis of physiological signals is often more meaningful when presented along with situational context awareness which needs to be embedded into the development of continuous monitoring and predictive systems to ensure its e effectiveness and robustness. Therefore, there is a need to develop improved and more comprehensive approaches towards studying interactions and correlations between multi modal clinical time series data. This is important because studies continue to show that humans are poor in reasoning about changes affecting more than two signals.

Genomics: The cost to sequence the human genome (encompassing 30,000 to 35,000 genes) is rapidly decreasing with the development of high throughput sequencing technology. With implications for current public health policies and delivery of care, analysing genome scale data for developing actionable recommendations in a timely manner is a significant challenge to the field of computational biology. Cost and time to deliver recommendations are crucial in a clinical setting. Initiatives tackling this complex problem include tracking of 100000 subjects over 20 to 30 years using the predictive, preventive, participatory and personalized health, refer to as P4, utilizing such high density data for exploration, discovery and clinical translation demands novel big data approaches and analytics. Despite the enormous expenditure consumed by the current healthcare systems, clinical outcomes remain suboptimal. A key factor attributing towards such inefficiencies is the inability to effectively gather, share and use information in a more comprehensive manner within the healthcare systems. This is an opportunity for big data analytics to play a more significant role in aiding the exploration and discovery process, improving the delivery of care, helping to design and plan healthcare policy, providing a means for comprehensively measuring and evaluating the complicated and convoluted data of healthcare. More importantly, adoption of insights gained from big data analytics has the potential to save lives, improve care delivery, expand access to healthcare, align pay with performance, and help curb the vexing growth of healthcare costs.

Medical Imaging

Medical imaging is the technique and process of creating visual representations of the interior of a body for clinical analysis and medical intervention, as well as visual representation of the function of some organs or tissues (physiology). Medical imaging seeks to reveal internal structures hidden by the skin and bones, as well as to diagnose and treat disease. Medical imaging also establishes a database of normal anatomy and physiology to make it possible to identify abnormalities. Although imaging of removed organs and tissues can be performed for medical reasons, such procedures are usually considered part of pathology instead of medical imaging.

Popular Imaging Modalities in Healthcare Domain

COMPUTED TOMOGRAPHY (CT)

Computed Tomography (CT), also commonly referred to as a CAT scan, is a medical imaging method that combines multiple X-ray projections taken from different angles to produce detailed cross-sectional images of areas inside the body. CT images allow doctors to get very precise, 3-D views of certain parts of the body, such as soft tissues, the pelvis, blood vessels, the lungs, the brain, the heart, abdomen and bones. CT is also often the preferred method of diagnosing many cancers, such as liver, lung and pancreatic cancers.

MAGNETIC RESONANCE IMAGING (MRI)

Magnetic Resonance Imaging (MRI) is a medical imaging technology that uses radio waves and a magnetic field to create detailed images of organs and tissues. MRI has proven to be highly effective in diagnosing a number of conditions by showing the difference between normal and diseased soft tissues of the body.

POSITRON EMISSION TOMOGRAPHY (PET)

Positron Emission Tomography (PET) is a nuclear imaging technique that provides physicians with information about how tissues and organs are functioning. PET, often used in combination with CT imaging, uses a scanner and a small amount of radiopharmaceuticals which is injected into a patient's vein to assist in making detailed, computerized pictures of areas inside the body.

ULTRASOUND

Diagnostic ultrasound, also known as medical sonography or ultrasonography, uses high frequency sound waves to create images of the inside of the body. The ultrasound machine sends sound waves into the body and is able to convert the returning sound echoes into a picture. Ultrasound technology can also produce audible sounds of blood flow, allowing medical professionals to use both sounds and visuals to assess a patient's health.

X-RAY

X-ray technology is the oldest and most commonly used form of medical imaging. X-rays use ionizing radiation to produce images of a person's internal structure by sending X-ray beams through the body, which are absorbed in different amounts depending on the density of the material. In addition, included as "x-ray type" devices are also mammography, interventional radiology, computed radiography, digital radiography and computed tomography (CT).

Big Data Analytics in Medical Imaging

Medical Image Processing from Big Data Point of View

Medical imaging provides important information on anatomy and organ function in addition to detecting diseases states. Moreover, it is utilized for organ delineation, identifying tumours, spinal deformity diagnosis, artery stenosis detection, aneurysm detection, etc. In these applications image processing techniques such as enhancement, segmentation and de-noising in addition to machine learning methods are employed. As the size and dimensionality of data increase, understanding the dependencies among the data and designing efficient, accurate and computationally effective methods demand new computer aided techniques and platforms. The rapid growth in the number of health care organizations as well as the number of patients has resulted in the greater use of computer aided medical diagnostics and decision support systems in clinical settings. Many areas in health care such as diagnosis, prognosis and screening can be improved by utilizing computational intelligence. The integration of computer analysis with appropriate care has potential to help clinicians improve diagnostic accuracy. The integration of medical images with other types of electronic health record (EHR) data and genomic data can improve the accuracy and reduce the time taken for a diagnosis.

Data Produced by Imaging Techniques

Medical imaging encompasses a wide spectrum of different image acquisition methodologies typically utilized for a variety of clinical applications. For example, visualizing blood vessel structure can be performed using magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and photoacoustic imaging. From a data dimension point of view, medical images might have 2, 3 and four dimensions. Positron emission tomography (PET), CT, 3D ultrasound and functional MRI (f MRI) are considered as multidimensional medical data. Modern medical image technologies can produce high resolution images such as respiration correlated or four dimensional computed tomography (4D CT). Higher resolution and dimensions of these images generates large volumes of data requiring high performance computing (HPC) and advanced analytical methods for its utilization. For instance, microscopic scans of a human brain with high resolution can require 66TB of storage space. Although the volume and variety of medical data make its analysis a big challenge, advances in medical imaging could make individualized care more practical and provide quantitative information in variety of applications such as disease strati cation, predictive modelling, decision making systems and so on. In the following we refer to two medical imaging techniques and one of their associated challenges.

Molecular imaging is a non-invasive technique of cellular and sub cellular events which has the potential for clinical diagnosis of disease states such as cancer. However, in order to make it clinically applicable for patients, the interaction of radiology, nuclear medicine and biology is crucial that could complicate its automated analysis.

Microwave imaging is an emerging methodology that could create a map of electromagnetic wave scattering arising from the contrast in the dielectric properties of different tissues. It has both functional and physiological information encoded in the dielectric properties which can help differentiate and characterize different tissues and/or pathologies. However, microwaves have scattering behaviour that makes retrieval of information a challenging task.

The integration of images from different modalities and/or other clinical and physiological information could improve the accuracy of diagnosis and outcome prediction of disease. Liebeskind and Feldman explored advances in neurovascular imaging and the role of

multimodal CT or MRI including angiography and perfusion imaging on evaluating the brain vascular disorder, and achieving precision medicine. Delayed enhanced MRI is used for exact assessment of myocardial infarction scar and electro anatomic mapping (EAM) can help in identifying the sub endocardial extension of infarct. The role of evaluating both MRI and CT images to increase the accuracy of diagnosis in detecting the presence of erosions and osteophytes in the temporomandibular joint (TMJ) has been investigated by Hussain et al. According to this study simultaneous evaluation of all the available imaging techniques is an unmet need.

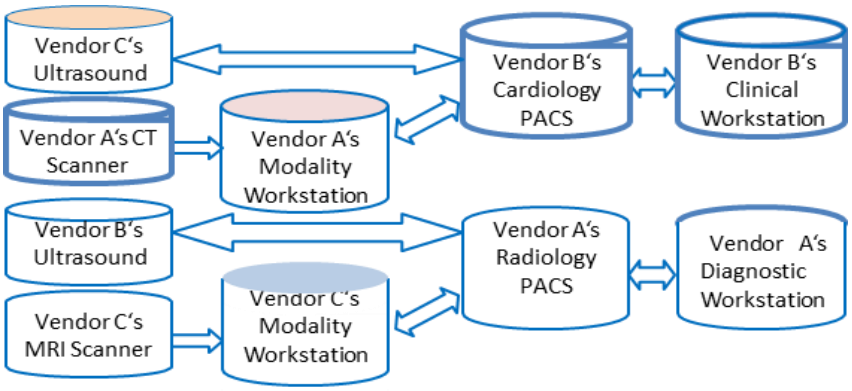
Advanced Multimodal Image Guided Operating (AMIGO) suite has been designed which has angiographic X ray system, MRI, 3D ultrasound and PET/CT imaging in the operating room. This system has been used for cancer therapy and showed the improvement in localization and targeting an individual's diseased tissue.

Besides the huge space required for storing all the data and their analysis, ending the map and dependencies among different data types are challenges for which there is no optimal solution yet.

Evolution of Medical Imaging Informatics

Radiology invented the concept of Picture Archival & Communication Systems (PACS) "Father of PACS‟ - The late Samuel J. Dwyer, III, PhD. It is a solution that is born out of real-world needs, due to a need to improve diagnostic capabilities. These needs are so effectively fulfilled that PACS these days are no longer limited to only medical images nor strictly for the radiology discipline PACS (next to the EMR) is to be one of the most significant clinical information systems in the healthcare enterprise.

Figure below shows the Typical Multi-PACS Environment:



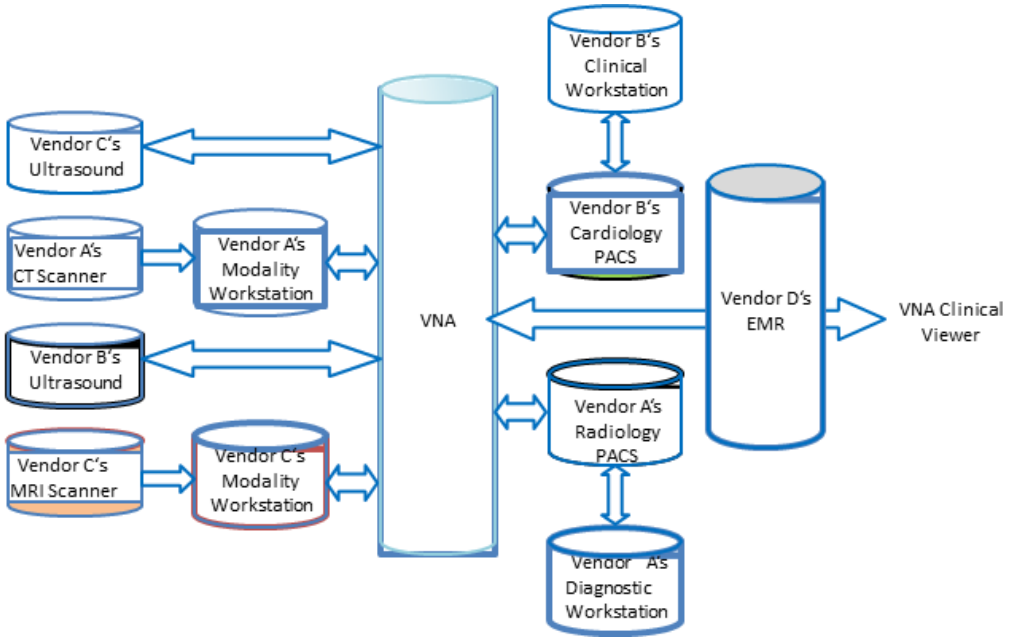
Typical Multi-PACS Environment

Figure 10

The “Traditional” implementation resulted in a “Silo model” Such PACS silos were appropriate for departmental implementations. As health IT adoption moves onto the enterprise levels, "Silo model” no longer serves clinical or operational needs. From a Technologist, Specialist, Clinician, Administrative, Technical and most importantly Patient's perspective. In the modern healthcare enterprise, we need a patient centric workflow.

The Problem associated with PACS leads to Vendor Neutral Archive (VNA) Architecture that Helps to rectify inherent issues with DICOM such as Private Tags, Transfer Syntax, Data Migration and Enables Single Web Client access across all [13-15].

Figure below shows the typical VNA Architecture:



VNA Architecture

Figure 11

The VNA Architecture has enabled solution architects "pushed" the limits a little further. towards the clouds Extending both the benefits of Cloud and VNA Architecture This actually makes perfect sense Vendor Neutrality for Interoperability Facilitating Patient Centric Care Cost by a need to use Gives new meaning to the phase "Imaging Anytime, Anywhere, Any place" [3].

The volume of medical images is growing exponentially. For instance, Image CLEF medical image dataset contained around 66,000 images between 2005 and 2007 while just in the year of 2013 around 300,000 images were stored every day. In addition to the growing volume of images, they differ in modality, resolution, dimension and quality which introduce new challenges such as data integration and mining specially if multiple datasets are involved. Compared to the volume of research that exists on single modal medical image analysis, there are considerably lesser number of research initiatives on multi modal image analysis.

When utilizing data at a local/institutional level, an important aspect of the research is on how the developed system is evaluated and validated. Having annotated data or a structured method to annotate new data is a real challenge. This becomes even more challenging when large scale data integration from multiple institutions are taken into account. For the same applications and the same modality such as CT scans for traumatic brain injury, different institutes might use different settings for image acquisitions. In order to bene t the multi modal images and their integration with other medical data, new analytical methods with real time feasibility and scalability are required. In the following we look at analytical methods that deal with some aspects of big data.

Analytical Methods

The goal of medical image analytics is to improve the interpretability of depicted contents. Many methods and frameworks have been developed for medical image processing. However, these methods are not necessarily applicable for big data analytics.

One of the frameworks developed for analysing and transformation of very large datasets is Hadoop that uses MapReduce. MapReduce is a programming paradigm that provides scalability across many servers in a Hadoop cluster with a broad variety of real world applications. However, it doesn't perform well with input output intensive tasks [4]. MapReduce framework has been used in to increase the speed of three large scale medical image processing use cases, (I) employing a well-known machine learning method, support vector machines (SVM), to end optimal parameter for lung texture classification (ii) content based medical image indexing, and (iii) wavelet analysis for solid texture classification. In this framework, a cluster of heterogeneous computing nodes with a maximum of 42 concurrent map tasks was set up and the speedup around 100 was achieved. In other words, total execution time for ending optimal SVM parameters was reduced from about 1000h to around 10h. Designing a fast method is crucial in some applications such as trauma assessment in critical care where the end goal is to utilize such imaging techniques and its analysis within what is considered as a golden hour of care. Therefore, execution time or real time feasibility of developed methods is of importance. Accuracy is another factor that should be considered in designing an analytical method. Finding dependencies among different types of data could help improve the accuracy. A hybrid machine learning method has been developed in that classifies schizophrenia patients and healthy controls using f MRI images and single nucleotide polymorphism (SNP) data. A classification accuracy of 87% has been achieved which is higher than using either data alone. Alfonso et al. have compared some organ segmentation methods when data is considered as big data. They have proposed a method that incorporates both the local contrast of the image and atlas probabilistic information. An average of 33% improvement has been achieved compared to using only atlas information. Tsymbal et al. have designed a clinical decision support system that exploits discriminative distance learning with significantly lower computational complexity compared to classical alternatives and hence this system is more scalable to retrieval from big data. A computer aided decision support system was developed by Wenan etal that can assist physicians to provide accurate treatment planning for patients suffering from traumatic brain injury (TBI). In this method, patient's demographic information, medical records, and features extracted from CT scans were combined to predict the level of intracranial pressure (ICP). The accuracy, sensitivity and specificity were reported to be around 70.3%, 65.2%, 73.7% respectively, molecular imaging and its impact on cancer detection and cancer drug improvement are discussed. The proposed technology is designed to aid in the early detection of cancer by integrating molecular and physiological information with anatomical information. Using this imaging technique for patients with advanced ovarian cancer, the accuracy of the predictor of response to a special treatment has been increased compared to other clinical or histopathologic criteria. A hybrid digital optical correlator (HDOC) has been designed to speed up the correlation of images. HDOC can be employed to compare images in the absence of coordinate matching or geo registration. In this multichannel correlator method, the computation is performed in the storage medium which is a volume holographic memory. These features could help HDOC to be applicable in the area of big data analytics.

In addition to developing analytical methods, e orts have been made for collecting, compressing, sharing and anonymizing medical data. One example is iDASH (integrating data for analysis, anonymization, and sharing) which is a centre for biomedical computing. It

focuses on algorithms and tools for sharing data in a privacy preserving manner. The goal of iDASH is to bring together a multi institutional team of quantitative scientists to develop algorithms and tools, services, and a biomedical cyber infrastructure to be used by biomedical and behavioural researchers. Another example of a similar approach is Healthy child consortium of 14 academics, industry, and clinical partners with the aim of developing an integrated healthcare platform for European Paediatrics.

Based on Hadoop platform, a system has been designed for exchanging, storing and sharing electronic medical records (EMR) among different healthcare systems. This system can also help users retrieve medical images from a database. Medical data has been investigated from an acquisition point of view where patients' vital data is collected through a network of sensors. This system delivers data to a cloud for storage, distribution and processing. A prototype system has been implemented in to handle standard store/query/retrieve requests on a database of Digital Imaging and Communications in Medicine (DICOM) images. This system uses Microsoft Windows Azure as a cloud computing platform [3].

When dealing with very large volume of data, compression techniques can help overcome data storage and network bandwidth limitations. Many methods have been developed for medical image compression. However, there are a few methods developed for big data compression. A method has been designed to compress both high throughput sequencing dataset and the data generated from calculation of log odds of probability error for each nucleotide while the maximum compression ratios of 400 and 5 haves been achieved respectively. This dataset has medical and biomedical data including genotyping, gene expression, proteomic measurements with demographics, laboratory values, images, therapeutic interventions, and clinical phenotypes for Kawasaki Disease(KD). By illustrating the data with a graph model, a framework for analysing large scale data has been presented. For this model, the fundamental signal processing techniques such as filtering and Fourier transform are implemented, the application of simplicity and power (SP) theory of intelligence in big data has been investigated. The goal of SP theory is to simplify and integrate concepts from multiple fields such as artificial intelligence, mainstream computing, mathematics, and human perception and cognition that can be observed as a brain-like system. The proposed SP sys-tem performs lossless compression through the matching and unification of patterns. However, this system is still in the design stage and cannot be supported by today's technologies [11].

There are some limitations in implementing the application specific compression methods on both general purpose processors and parallel processors such as graphics processing units (GPUs) as these algorithms need highly variable control and complex bit manipulations which are not well suited to GPUs and pipeline architectures. To overcome this limitation, an FPGA implementation is proposed for LZ factorization which decreases the computational burden of the compression algorithm. A lossy image compression has been introduced that reshapes the image in such a way that if the image is uniformly sampled, sharp features have a higher sampling density than the coarse ones. This method is claimed to be applicable for big data compression. However, for medical applications lossy methods are not applicable in most cases as fidelity is important and information must be preserved.

Table 1: Challenges facing medical image analysis

Challenges	Description and Possible Solutions
Pre-processing	Medical images suffer from different types of noise/artefacts and missing data. Noise reduction, artefact removal, missing data handling, contrast adjusting and etc could enhance the quality of images an Employing multimodal data could be beneficial for this purposed increase the performance of processing methods.
Compression	Reducing the volume of data while maintaining important data such as anatomically relevant data
Parallelization/ Real-time realization	Developing scalable/parallel methods and frameworks to speed up the analysis/processing
Registration	Aligning consecutive slices/frames from one scan or corresponding images from different modalities
Sharing /Security/ Anonymization	Integrity, privacy and confidentiality of data must be protected
Segmentation	Delineation of anatomical structure such as vessels, bones
Data Integration/Mining	Finding dependencies/patterns among multimodal data and/or the data captured at different time points in order to increase the accuracy of diagnosis, prediction and overall performance of the system
Validation	Assessing the performance or accuracy of the system/method. Validation can be objective or subjective. For the former, annotated data is usually required

These techniques are among a few techniques that have been either designed as prototypes or developed with limited applications. Developing methods for processing/analysing a broad range and large volume of data with acceptable accuracy and speed is still critical. In Table 1, we summarize the challenges facing medical image processing. When dealing with big-data, these challenges seemed to be more serious and on the other hand analytical methods could bene t the big data to handle them.

Conclusion

Big data analytics which leverages legions of disparate, structured and unstructured data sources is going to play a vital role in how healthcare is practiced in the future. One can already see a spectrum of analytics being utilized, aiding in the decision making and performance of healthcare personnel and patients. Here we focused on three areas of interest: medical image analysis, physiological signal processing and integration of physiological data with genomic data. The exponential growth of the volume of medical images forces computational scientists to come up with innovative solutions to process this large volume of data in tractable timescales. The trend of adoption of computational systems for physiological signal processing from both research and practicing medical professionals is growing steadily with the development of some very imaginative and incredible systems that help save lives. Developing a detailed model of a human being by combining physiological data and high throughput ` omics' techniques has the potential to enhance our knowledge of disease states and help in the development of blood based diagnostic tools. Medical image analysis, signal processing of physiological data, and integration of physiological and ` omics' data face similar challenges and opportunities in dealing with disparate structured and unstructured big data sources.

Medical image analysis covers many areas such as image acquisition, formation/reconstruction, enhancement, transmission, and compression. New technological advances have resulted in higher resolution, dimension and availability of multi modal images which lead to the increase in accuracy of diagnosis and improvement of treatment. However, integrating medical images with different modalities or with other medical data is a potential opportunity. New analytical frameworks and methods are required to analyze these data in a clinical setting. These methods address some concerns, opportunities and challenges such as: features from images which can improve the accuracy of diagnosis, ability to utilize disparate sources of data to increase the accuracy of diagnosis and reducing cost, and improving the accuracy of processing methods such as medical image enhancement, registration and segmentation to deliver better recommendations at the clinical level.

References

1. Jeffrey Dean and Sanjay Ghemawat. “Mapreduce: simplified data processing on large clusters”, Communications *of the ACM,* vol. 51, no. 1, pp.107-113, 2008.
2. Marco Viceconti, Peter Hunter, and Rod Hose “Big Data, Big Knowledge: Big Data for Personalized Healthcare”, IEEE Journal of Biomedical And Health Informatics, vol. 19, no. 4, July 2015.
3. Alberto Bartesaghi, Guillermo Sapiro, and Sriram Subramaniam “An Energy-Based Three-Dimensional Segmentation Approach for the Quantitative Interpretation of Electron Tomograms”, IEEE Transactions on Image Processing, vol. 14, no. 9, Sept. 2005.
4. Chao-Tung Yang, Lung-Teng Chen, Wei-Li Chou, and Kuan-Chieh Wang. “Implementation of a medical image file accessing system on cloud computing”, *IEEE International Conference on Computational Science and Engineering (CSE),* pages 321-326. IEEE, 2010.
5. Carlos O R, Fernando L K, Carlos B W, Jorge W, Armando F, and Giovanni S S. “A cloud computing solution for patient's data collection in health care institutions”, *IEEE International Conference on eHealth, Telemedicine, and Social Medicine (ETELEMED),* pp 95-99, 2010.
6. Chia C T, Jonathan M, Christopher W, Alex S, Cesar D, David H, and Travis N. A “medical image archive solution in the cloud”, *IEEE International Conference on Software Engineering and Service Sciences,* pp. 431-434, 2010.
7. Alexey Tsymbal, Eugen Meissner, Michael Kelm, and Martin Kramer. “Towards cloud-based image-integrated similarity search in big data”, *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI),* pp. 593-596. IEEE, 2014.
8. Wenan Chen, Charles Cockrell, KR Ward, and Kayvan Najarian. “Intracranial pressure level prediction in traumatic brain injury by extracting features from multiple sources and using machine learning methods”, *IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* pp. 510-515, 2010.
9. https://cloudera.com
10. https://hadoop.apache.
11. Antoine Widmer, Roger Schaer, Dimitrios Markonis, and Henning Müller. “Gesture interaction for content based medical image retrieval”, *Proceedings of International Conference on Multimedia Retrieval, ACM,* pp. 503-506, 2014.
12. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. “The hadoop distributed file system” *IEEE Symposium on Mass Storage Systems and Technologies (MSST),* pp. 1-10, 2010.
13. Dalia Sobhy, Yasser El-Sonbaty, and M Abou Elnasr. “Medcloud: healthcare cloud computing system”, *IEEE International Conference on Internet Technology And Secured Transactions,* pp. 161-166, 2012.

14. Akgül, Ceyhun Burak, et al. "Content-based image retrieval in radiology: current status and future directions." *Journal of Digital Imaging,* vol. 24 No. 2 pp. 208-222, 2011.

15. Müller, Henning, et al. "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions", *International Journal of Medical Informatics,* vol. 73 no.1 pp. 1-24, 2004.