Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.675 – 688

RESEARCH ARTICLE

MapReduce Framework Implementation on the Prescriptive Analytics of Health Industry

Lalit Malik
Student, M.Tech, SBMNEC, Rohtak
Malik.kabir22@gmail.com

Sunita Sangwan
Prof, CSE Dept, SBMNEC, Rohtak

ABSTRACT: Prescriptive analytics is a type of business analytics that focuses on finding the best course of action for a given situation, and belongs to a portfolio of analytic capabilities that include descriptive and predictive analytics.

Keywords: Mapreduce, HDFS, DFS, Hive, Pig, Hadoop

INTRODUCTION

“90% of the world’s data was generated in the last few years.”

Relevant literature cited in this paper related to “MapReduce, Hadoop, clinical data, and biomedical/bioinformatics applications of MapReduce” was obtained from PubMed, IEEEXplore, Springer, and BioMed Central databases. The MapReduce programming framework was first introduced to industry in 2006. And thus the literature search concentrated on 2007 to 2014. A total of 32 articles were found based on the use of the MapReduce framework to process the clinical big data and its application using the Hadoop platform.

In this review we start by listing the different types of big health major datasets, followed by the efforts that are developed to leverage the data for analytical advantages. These advantages are mainly focused on descriptive and predictive analytics. The major reason for using the MapReduce programming framework in the reviewed efforts is to speed up these kind of analytics. This is due the fact that these kinds of analytic algorithms are very well developed and tested for the MapReduce framework and the Hadoop platform can handle a huge amount of data in a small amount of time. The prescriptive analytics require data sharing among computing nodes, which unfortunately cannot be achieved easily (i.e. sophisticated programs with a great deal of data management) using MapReduce, and thus, not all optimization problems (i.e. prescriptive analytics) can be implemented on the MapReduce framework.

The review section is followed by a challenges and future trends section that highlights the use of the Map Reduce programming framework and its open source implementation Hadoop for processing clinical big data. This is followed by our perspective and use cases on how to leverage clinical big data for novel analytics.

Clinical big data analysis

The exponential production of data in recent years has introduced a new area in the field of information technology known as 'Big Data'. In a clinical setting such datasets are emerging from large-scale laboratory information system (LIS) data, test utilization data, electronic medical record (EMR), biomedical data, biometrics data, gene expression data, and in other areas. Massive datasets are extremely difficult to analyse and query using traditional mechanisms, especially when the queries themselves are quite complicated. In effect, a MapReduce algorithm maps both the query and the dataset into constituent parts. The mapped components of the query can be processed simultaneously – or reduced – to rapidly return results.

Big datasets of clinical, biomedical, and biometric data have been processed successfully using the MapReduce framework on top of the Hadoop distributed file system. An overview of the Hadoop platform, MapReduce framework and its current applications has been reported for the field of bioinformatics. The promise of big data analytics in bioinformatics and health care in general has previously been described . However our review enlarges the scope to the application of the MapReduce framework and its open source implementation Hadoop to a wide range of clinical big data including:

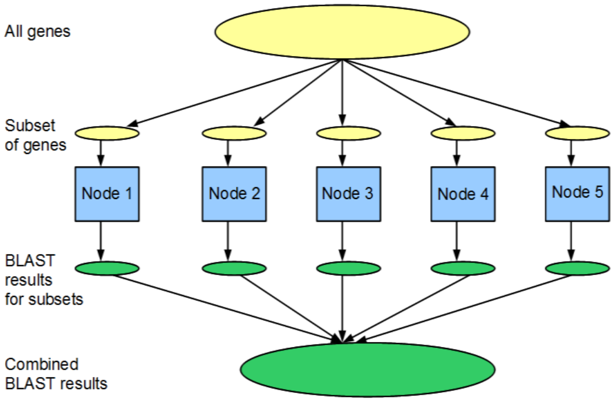


Figure 1 : Big data block diagram

1. Publicly available clinical datasets: online published datasets and reports from the United States Food and Drug Administration (FDA).

2. Biometrics datasets: containing measurable features related to human characteristics. Biometrics data is used as a form of identification and access control .

3. Bioinformatics datasets: biological data of a patient (e.g. protein structure, DNA sequence, etc.).

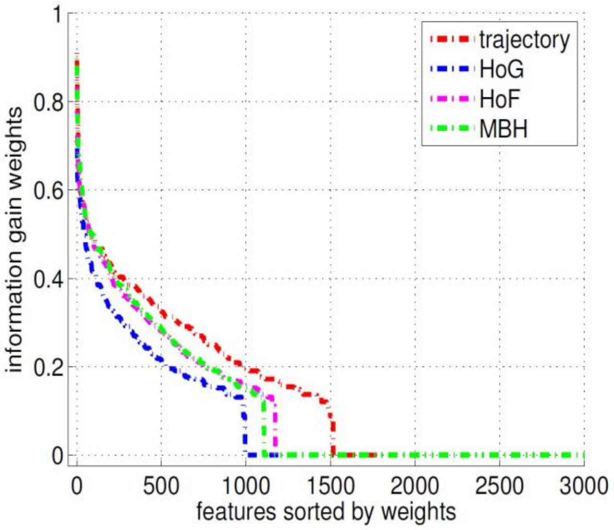
4. Biomedical signal datasets: data resulting from the recording of vital signs of a patient (e.g. electrocardiography (ECG), electroencephalography (EEG), etc.).

5. Biomedical image datasets: data resulting from the scanning of medical images (e.g. ultrasound imaging, magnetic resonance imaging (MRI), histology images, etc.). Moreover, our review presents a detailed discussion about the various types of clinical big data, challenges and consequences relevant to the application of big data analytics in a health care facility. This review is concluded with the future potential applications of the MapReduce programming framework and the Hadoop platform applied to clinical big data.

. Public databases

A Map Reduce-based algorithm has been proposed for common adverse drug event (ADE) detection and has been tested in mining spontaneous ADE reports from the United States FDA. The purpose of this algorithm was to investigate the possibility of using the Map Reduce framework to speed up biomedical data mining tasks using this pharma covigilance case as one specific example. The results demonstrated that the MapReduce programming framework could improve the performance of common signal detection algorithms for pharmacovigilance in a distributed computation environment at approximately linear speedup rates. The MapReduce distributed architecture and high dimensionality compression via Markov boundary feature selection have been used to identify unproven cancer treatments on the World Wide Web. This study showed that unproven treatments used distinct language to market their claims and this language was learnable, and through distributed parallelization and state of the art feature selection , it is possible to build and apply models with large scalability.

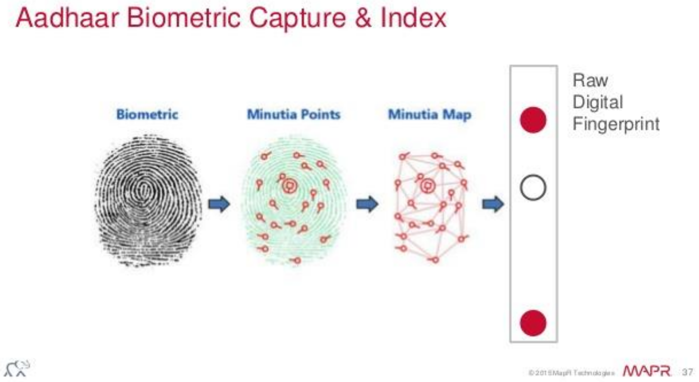
A novel system known as GroupFilterFormat has been developed to handle the definition of field content based on a Pig Latin script . Dummy discharge summary data for 2.3 million inpatients and medical activity log data for 950 million events were processed. The response time was significantly reduced and a linear relationship was observed between the quantity of data and processing time in both a small and a very large dataset. The results show that doubling the number of nodes resulted in a 47% decrease in processing time.



Biometrics

The MapReduce programming framework has also been used to classify biometric measurements using the Hadoop platform for face matching, iris recognition, and fingerprint recognition. A biometrics prototype system has been implemented for generalized searching of cloud-scale biometric data and matching a collection of synthetic human iris images. A biometric-capture mobile phone application has been developed for secure access to the cloud. The biometric capture and recognition are performed during a standard Web session. The Hadoop platform is used to establish the connection between a mobile user and the server in the cloud.

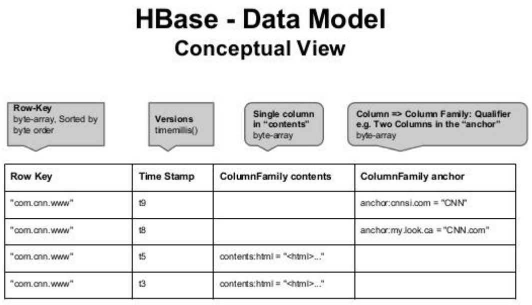
.



Aadhaar Biometric Capture & Index

Biomedical signal analysis

The parallel ensemble empirical mode decomposition (EEMD) algorithm has been implemented on top of the Hadoop platform in a modern cyber infrastructure . The algorithm described a parallel neural signal processing with EEMD using the Map Reduce framework. Test results and performance evaluation show that parallel EEMD can significantly improve the performance of neural signal processing. A novel approach has been proposed to store and process clinical signals based on the Apache HBase distributed column-store and the Map Reduce programming framework with an integrated Web-based data visualization layer.



HBase - Data Model

Conceptual View

Row-Key byte-array, Sorted by byte order
Versions timemillis()
Single column in "contents" byte-array
Column => Column Family: Qualifier e.g. Two Columns in the "anchor" byte-array

Row Key	Time Stamp	ColumnFamily contents	ColumnFamily anchor
"com.cnn.www"	t9		anchor:cnnsi.com = "CNN"
"com.cnn.www"	t8		anchor:my.look.ca = "CNN.com"
"com.cnn.www"	t5	contents:html = "<html>..."	
"com.cnn.www"	t3	contents:html = "<html>..."	

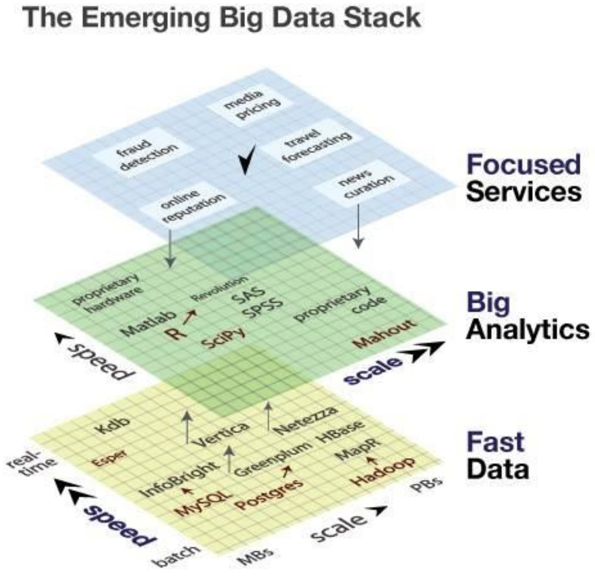
Biomedical image analysis

The growth in the volume of medical images produced on a daily basis in modern hospitals has forced a move away from traditional medical image analysis and indexing approaches towards scalable solutions . Map Reduce has been used to speed up and make possible three large–scale medical image processing use–cases: (1) parameter optimization for lung texture classification using support vector machines (SVM), (2) content–based medical image indexing/retrieval, and (3) dimensional directional wavelet analysis for solid texture classification . A cluster of heterogeneous computing nodes was set up using the Hadoop platform allowing for a maximum of 42 concurrent map tasks. The majority of the machines used were desktop computers that are also used for regular office work. The three use–cases reflect the various challenges of processing medical images in different clinical scenarios.

An ultrafast and scalable cone-beam computed tomography (CT) reconstruction algorithm using Map Reduce in a cloud-computing environment has been proposed . The algorithm accelerates the Feldcamp-Davis-Kress (FDK) algorithm by porting it to a MapReduce implementation. The map functions were used to filter and back-project subsets of projections, and reduce functions to aggregate that partial back-projection into the whole volume. The speed up of reconstruction time was found to be roughly linear with the number of nodes employed.

Tabular data includes a summary of the discussed literature on clinical big data analysis using the

Map Reduce programming framework. It tabulates the studies referenced in this paper grouped by relevant categories to indicate the following fields: study name, year, and technology used, and potential application of the algorithm or the technology used



Challenges and future trends

Health care systems in general suffer unsustainable costs and lack data utilization .

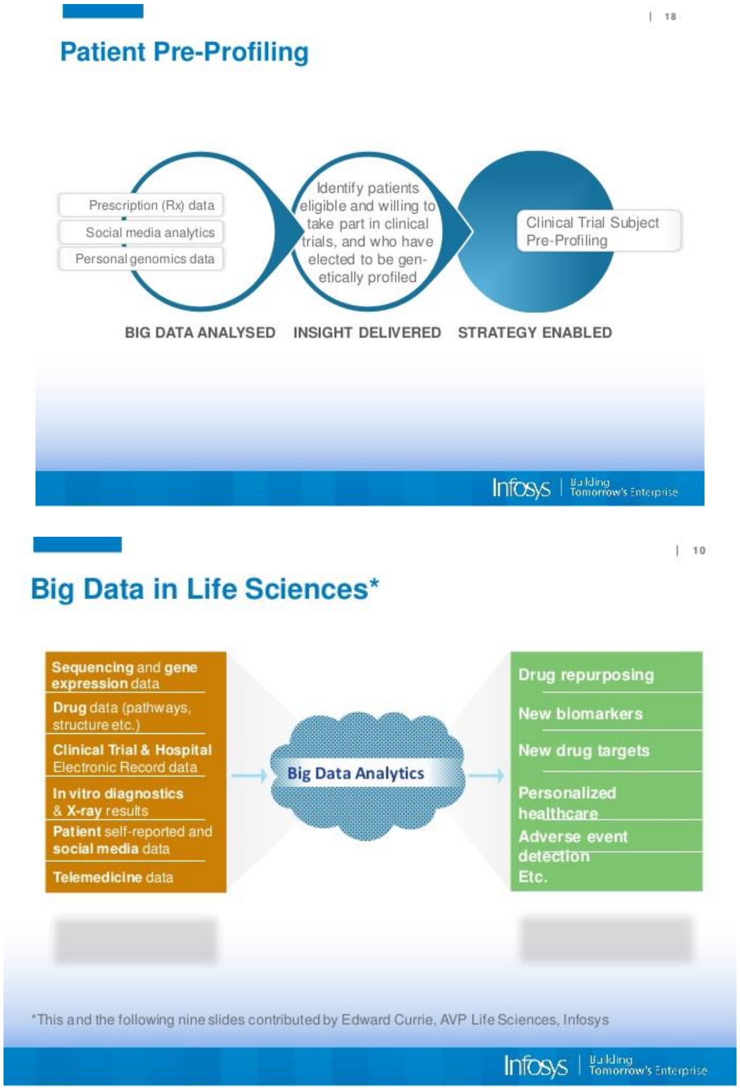
Therefore there is a pressing need to find solutions that can reduce unnecessary costs. Advances in health quality outcomes and cost control measures depend on using the power of large integrated databases to underline patterns and insights. However, there is much less certainty on how this clinical data should be collected, maintained, disclosed, and used. The problem in health care systems is not the lack of data, it is the lack of information that can be utilized to support critical decision-making . This presents the following challenges to big data solutions in clinical facilities:

1- Technology straggling. Health care is resistant to redesigning processes and approving technology that influences the health care system .

2- Data dispersion. Clinical data is generated from many sources (e.g. providers, labs, data vendors, financial, regulations, etc.) this motivates the need for data integration and maintaining mechanism to hold the data into a flexible data warehouse.

3- Security concerns and privacy issues. There are lots of benefits from sharing clinical big data between researchers and scholars, however these benefits are constricted due to the privacy issues and laws that regulate clinical data privacy and access .

4- Standards and regulations. Big data solution architectures have to be flexible and adoptable to manage the variety of dispersed sources and the growth of standards and regulations (e.g. new encryption standards that may require system architecture modifications) that are used to interchange and maintain data.



| 18

Patient Pre-Profiling

Prescription (Rx) data

Social media analytics

Personal genomics data

Identify patients eligible and willing to take part in clinical trials, and who have elected to be genetically profiled

Clinical Trial Subject Pre-Profiling

BIG DATA ANALYSED

INSIGHT DELIVERED

STRATEGY ENABLED

Infosys | Building Tomorrow's Enterprise

| 10

Big Data in Life Sciences*

Sequencing and **gene expression** data

Drug data (pathways, structure etc.)

Clinical Trial & Hospital Electronic Record data

In vitro diagnostics & **X-ray** results

Patient self-reported and **social media** data

Telemedicine data

Big Data Analytics

Drug repurposing

New biomarkers

New drug targets

Personalized healthcare

Adverse event detection

Etc.

*This and the following nine slides contributed by Edward Currie, AVP Life Sciences, Infosys

Infosys | Building Tomorrow's Enterprise

Infosys framework for various medical stream data analytics

An outlook for the future

Big Data has a substantial potential to unlock the whole health care value chain . Big data analytics changed the traditional perspective of health care systems from finding new drugs to patient-central health care for better clinical outcomes and increased efficiency. The future applications of big data in the health care system have the potential of enhancing and accelerating interactions among clinicians, administrators, lab directors, logistic mangers, and researchers by saving costs, creating better efficiencies based on outcome comparison, reducing risks, and improving personalized care.

The following is a list is of potential future applications associated with clinical big data.

1- E-clinics, E-medicine, and similar case retrieval applications based on text analytics applications.

Large amounts of health data is unstructured as documents, images, clinical or transcribed notes. Research articles, review articles, clinical references, and practice guidelines are rich sources for text analytics applications that aim to discover knowledge by mining these type of text data.

2- Genotyping applications.

Genomic data represent significant amounts of gene sequencing data and applications are required to analysis and understand the sequence in regards to better understanding of patient treatment.

3- Mining and analysis of biosensors applications.

Streamed data home monitoring, tele-health, handheld and sensor-based wireless are well established data sources for clinical data.

4- Social media analytics applications.

Social media will increase the communication between patients, physician and communities. Consequently, analytics are required to analyse this data to underline emerging outbreak of disease, patient satisfaction, and compliance of patient to clinical regulations and treatments.

5- Business and organizational modelling applications.

Administrative data such as billing, scheduling, and other non-health data present an exponentially growing source of data. Analysing and optimizing this kind of data can save large amounts of money and increase the sustainability of a health care facility.

The aforementioned types of clinical data sources provide a rich environment for research and give rise to many future applications that can be analysed for better patient treatment outcomes and a more sustainable health care system.

Clinical big data and the upcoming challenges

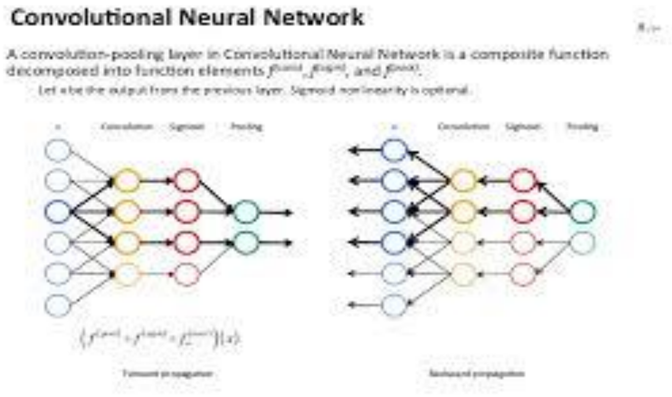
Big data by itself usually confers little direct advantage, however analytics based on big data can reveal many actionable insights that may prove useful in a clinical environment. This section describe the potential benefits and highlight potential application to leverage the clinical big data for analytical advantages using the Map Reduce programming framework and the Hadoop platform.

Epilepsy affects nearly 70 Million people around the world, and is categorized by the incident of extemporaneous seizures. Many medications can be given at high doses to inhibit seizures, however patients often suffer side effects. Even after surgical removal of epilepsy foci, many patients suffer extemporaneous seizures. Seizure prediction systems have the potential to help patients alleviate epilepsy episodes. Computational algorithms must consistently predict periods of increased probability of seizure incidence. If the seizure states can be predicted and classified using data mining algorithms, implementation of these algorithms on wearable devices can warn patients of impending seizures. Patients could avoid potentially unsuitable activities in potential seizures episode (e.g. driving and swimming). Seizure patterns are wide and complex resulting in a massive datasets when digitally acquired. MapReduce and Hadoop can be consciously used to train detection and forecasting models. Simulation of different concurrently seizures pattern require the development of complex distributed algorithms to deal with the massive datasets.

Understanding how the human brain functions is the main goal in neuroscience research . Non-invasive functional neuroimaging techniques, such as magneto encephalography (MEG) , can capture huge time series of brain data activities. Analysis of concurrent brain activities can reveal the relation between the pattern of recorded signal and the category of the stimulus and may provide insights about the brain functional foci (e.g. epilepsy, Alzheimer’s disease , and other neuro-pathologies, etc.). Among the approaches to analyse the relation between brain activity and stimuli, the one based on predicting the stimulus from the concurrent brain recording is called brain decoding.

The brain contains nearly 100 billion neurons with an average of 7000 synaptic connections each Tracing the neuron connections of the brain is therefore a tedious process due to the resulting massive datasets. Traditional neurons visualization methods cannot scale up to very large scale neuron networks. MapReduce framework and Hadoop platform can be used to visualize and recover neural network structures from neural activity patterns.

More than 44.7 million individuals in the United States are admitted to hospitals each year . Studies have concluded that in 2006 well over $30 billion was spent on unnecessary hospital admissions .To achieve the goal of developing novel algorithms that utilize patient data claim to predict and prevent unnecessary hospitalizations. Claims data analytics require text analytics, prediction and estimation models. The models must be tuned to alleviate the potential risk of decline the admission of patients who need to be hospitalized. This type of analysis is one application of fraud analysis in medicine.



Convolutional Neural Network

A convolution-pooling layer in Convolutional Neural Network is a composite function decomposed into function elements $f^{(conv)}$, $f^{(sigm)}$, and $f^{(pool)}$.

Let x be the output from the previous layer. Sigmoid nonlinearity is optional.

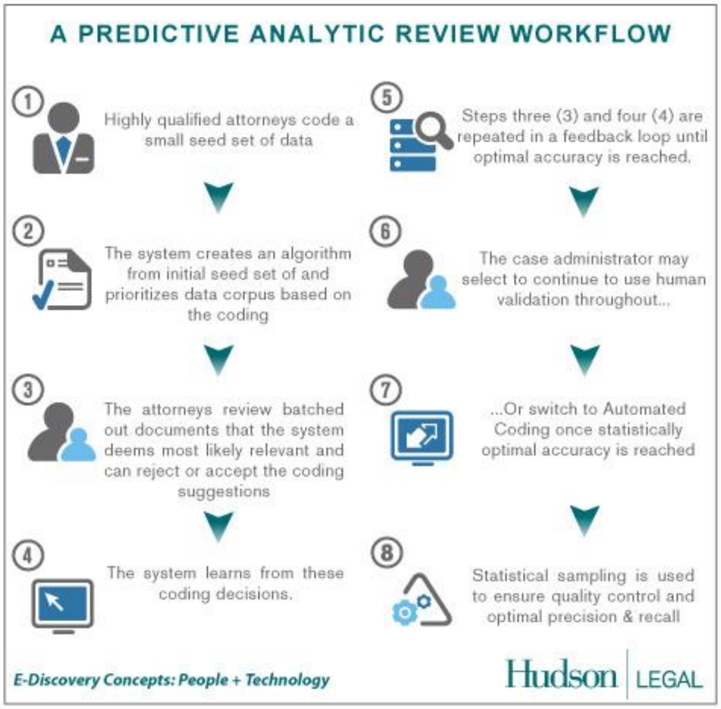
Conclusions

An integrated solution eliminates the need to move data into and out of the storage system while parallelizing the computation, a problem that is becoming more important due to increasing numbers of sensors and resulting data. And, thus, efficient processing of clinical data is a vital step towards multivariate analysis of the data in order to develop a better understanding of a patient clinical status (i.e. descriptive and predictive analysis). This highly demonstrates the significance of using the Map Reduce programming model on top of the Hadoop distributed processing platform to process the large volume of clinical data.

Big data solutions presents an evolution of clinical big data analysis necessitated by the emergence of ultra-large-scale datasets. Recent developments in open source software, that is, the Hadoop project and the associated software projects, provide a backbone foundation for scaling to terabytes and peta bytes data warehouses on Linux clusters, providing fault-tolerant parallelized analysis on such data using a programming framework named Map Reduce.

The Hadoop platform and the Map Reduce programming framework already have a substantial base in the bioinformatics community, especially in the field of next-generation sequencing analysis, and such use is increasing. This is due to the cost-effectiveness of the Hadoop-based analysis on commodity Linux clusters, and in the cloud via data upload to cloud vendors who have implemented Hadoop/HBase; and due to the effectiveness and ease-of-use of the Map Reduce method in parallelization of many data analysis algorithms.

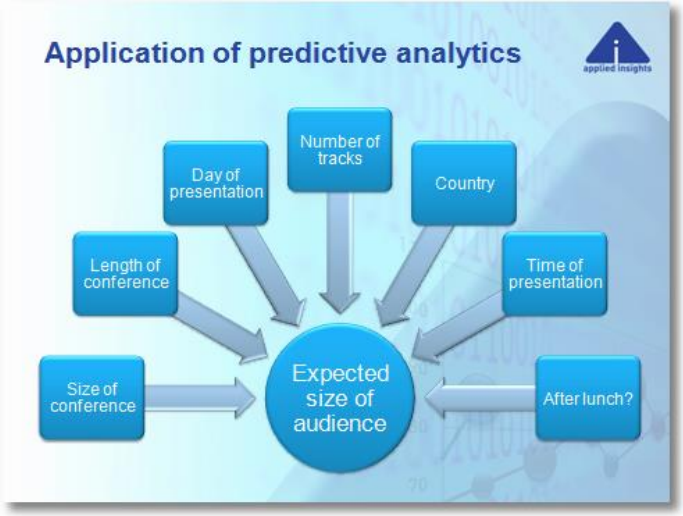
HDFS supports multiple reads and one write of the data. The write process can therefore only append data (i.e. it cannot modify existing data within the file). HDFS does not provide an index mechanism, which means that it is best suited to read-only applications that need to scan and read the complete contents of a file (i.e. MapReduce programs). The actual location of the data within an HDFS file is transparent to applications and external software. And, thus, Software built on top of HDFS has little control over data placement or knowledge of data location, which can make it difficult to optimize performance.



A PREDICTIVE ANALYTIC REVIEW WORKFLOW

1. Highly qualified attorneys code a small seed set of data
2. The system creates an algorithm from initial seed set of and prioritizes data corpus based on the coding
3. The attorneys review batched out documents that the system deems most likely relevant and can reject or accept the coding suggestions
4. The system learns from these coding decisions.
5. Steps three (3) and four (4) are repeated in a feedback loop until optimal accuracy is reached.
6. The case administrator may select to continue to use human validation throughout...
7. ...Or switch to Automated Coding once statistically optimal accuracy is reached
8. Statistical sampling is used to ensure quality control and optimal precision & recall

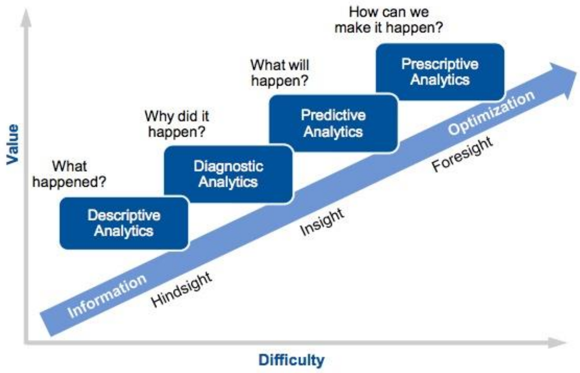
Future work on big clinical data analytics should emphasize modelling of whole interacting processes in a clinical setting (e.g. clinical test utilization pattern, test procedures, specimen collection/handling, etc.). This indeed can be constructed using inexpensive clusters of commodity hardware and the appropriate open source tool (e.g. HBase, Hive, and Pig Latin see for Hadoop related projects/ecosystems description and definition) to construct convenient processing tools for massive clinical data. These tools will form the basis of future laboratory informatics applications as laboratory data are increasingly integrated and consolidated.



Application of predictive analytics

Descriptive and predictive analysis

Figure 2. Gartner Analytic Ascendancy Model



Source: Gartner (March 2012)

References

1. Shuman S. Structure, mechanism, and evolution of the mRNA capping apparatus. Prog Nucleic Acid Res Mol Biol. 2000;66:1–40. [PubMed]
2. Rajaraman A, Ullman JD. Mining of Massive Datasets. Cambridge – United Kingdom: Cambridge University Press; 2012.
3. Coulouris GF, Dollimore J, Kindberg T. Distributed Systems: Concepts and Design: Pearson Education. 2005.
4. de Oliveira Branco M. Distributed Data Management for Large Scale Applications. Southampton – United Kingdom: University of Southampton; 2009.
5. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inform Sci Syst. 2014;2(1):3. doi: 10.1186/2047-2501-2-3. [Cross Ref]

6. BELL DE, Raiffa H, Tversky A. Descriptive, normative, and prescriptive interactions in decision making. Decis Mak. 1988;1:9–32.
7. Foster I, Kesselman C. The Grid 2: Blueprint for a new Computing Infrastructure. Houston – USA: Elsevier; 2003.

8. Owens JD, Houston M, Luebke D, Green S, Stone JE, Phillips JC. GPU computing. Proc IEEE.2008;96(5):879–899.
9. Satish N, Harris M, Garland M. Designing efficient sorting algorithms for manycore GPUs. Parallel & Distributed Processing, 2009 IPDPS 2009 IEEE International Symposium on: 2009. 2009. pp. 1–10. (IEEE).
10. He B, Fang W, Luo Q, Govindaraju NK, Wang T. Mars: a MapReduce framework on graphics processors. Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques: 2008. 2008. pp. 260–269. (ACM).
11. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM.2008;51(1):107–113. doi: 10.1145/1327452.1327492. [Cross Ref]
12. Peyton Jones SL. The Implementation of Functional Programming Languages (Prentice-Hall International Series in Computer Science) New Jersey – USA: Prentice-Hall, Inc; 1987.
13. Bryant RE. Data-intensive supercomputing: The case for DISC. Pittsburgh, PA – USA: School of Computer Science, Carnegie Mellon University; 2007. pp. 1–20.
14. White T. Hadoop: The Definitive Guide. Sebastopol – USA: " O'Reilly Media, Inc."; 2012.
15. Shvachko K, Kuang H, Radia S, Chansler R. The hadoop distributed file system. Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on: 2010. 2010. pp. 1–10. (IEEE).
16. The Apache Software Foundation. [http://apache.org/]
17. Olson M. Hadoop: Scalable, flexible data storage and analysis. IQT Quart. 2010;1(3):14–18.
18. Xiaojing J. Google Cloud Computing Platform Technology Architecture and the Impact of Its Cost. 2010 Second WRI World Congress on Software Engineering: 2010. 2010. pp. 17–20.
19. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Anthony S, Liu H, Wyckoff P, Murthy R. Hive: a warehousing solution over a map-reduce framework. Proc VLDB Endowment. 2009;2(2):1626–1629. doi: 10.14778/1687553.1687609. [Cross Ref]
20. Olston C, Reed B, Srivastava U, Kumar R, Tomkins A. Pig latin: a not-so-foreign language for data processing. Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data: 2008. 2008. pp. 1099–1110. (ACM).
21. The Platform for Big Data and the Leading Solution for Apache Hadoop in the Enterprise - Cloudera. [http://www.cloudera.com/content/cloudera/en/home.html]
22. DataStax. [http://www.datastax.com/]
23. Hortonworks. [http://hortonworks.com/]
24. MAPR. [http://www.mapr.com/products/m3]
25. Top 14 Hadoop Technology Companies. [http://www.technavio.com/blog/top-14-hadoop-technology-companies]
26. Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics. 2010;11(Suppl 12):S1. doi: 10.1186/1471-2105-11-S12-S1. [PMC free article] [PubMed] [Cross Ref]
27. Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct. 2012;7(1):43. doi: 10.1186/1745-6150-7-43. [PMC free article] [PubMed] [Cross Ref]
28. Microsoft Excel 2013: Spreadsheet software. [http://office.microsoft.com/en-ca/excel/]
29. Jonas M, Solangasenathirajan S, Hett D. Annual Update in Intensive Care and Emergency Medicine 2014. New York – USA: Springer; 2014. Patient Identification, A Review of the Use of Biometrics in the ICU; pp. 679–688.
30. Wang W, Haerian K, Salmasian H, Harpaz R, Chase H, Friedman C. AMIA Annual Symposium Proceedings: 2011. Bethesda, Maryland – USA: American Medical Informatics Association; 2011. A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from PubMed citations; p. 1464. [PMC free article] [PubMed]
31. Aphinyanaphongs Y, Fu LD, Aliferis CF. Identifying unproven cancer treatments on the health web: addressing accuracy, generalizability and scalability. Stud Health Technol Inform.2012;192:667–671. [PMC free article] [PubMed]
32. Yaramakala S, Margaritis D. Speculative Markov blanket discovery for optimal feature selection.Data Mining, Fifth IEEE International Conference on: 2005. 2005. p. 4. (IEEE).

33. Horiguchi H, Yasunaga H, Hashimoto H, Ohe K. A user-friendly tool to transform large scale administrative data into wide table format using a mapreduce program with a pig latin based script.BMC Med Inform Decis Mak. 2012;12(1):151. doi: 10.1186/1472-6947-12-151. [PMC free article][PubMed] [Cross Ref]
34. Kohlwey E, Sussman A, Trost J, Maurer A. Leveraging the cloud for big data biometrics: Meeting the performance requirements of the next generation biometric systems. Services (SERVICES), 2011 IEEE World Congress on: 2011. 2011. pp. 597–601. (IEEE).
35. Raghava N. Iris recognition on hadoop: A biometrics system implementation on cloud computing.Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on: 2011.2011. pp. 482–485. (IEEE).
36. Omri F, Hamila R, Foufou S, Jarraya M. Networked Digital Technologies. New York – USA: Springer; 2012. Cloud-Ready Biometric System for Mobile Security Access; pp. 192–200.
37. Chen W-P, Hung C-L, Tsai S-JJ, Lin Y-L. Novel and efficient tag SNPs selection algorithms.Biomed Mater Eng. 2014;24(1):1383–1389. [PubMed]
38. Zhang K, Sun F, Waterman MS, Chen T. Dynamic programming algorithms for haplotype block partitioning: applications to human chromosome 21 haplotype data. Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology: 2003. 2003. pp. 332–340. (ACM).