Use of Multi-Task Extreme Learning in Big Data for Medical Insights using Distributed Processing

R.V.Aswiga

Assistant Professor, Department of IT,Chennai-602105, India;aswiga91@gmail.com

Abstract

Bigdata is the context that collects lots of data from diverse sources, every second from different domains hence called as big data. But the model for generating data has been changed. In olden decades, few companies were generating data but all others are consuming those data. Whereas in present model, all of us are generating data and all of us are consuming data. Big data therefore refers to the tendency to make use of ever growing volumes of data. The past decade has witnessed the volatile growth of data for exemption. Large internet companies consistently produce hundreds of tera-bytes of logs and functional reports. Map reduce has proven itself to be an effectual tool for parallel data processing to resolve data skew problem. Imbalance in the total amount of work is often referred to as skew. The confront here is how to discover the partition breakpoints in a huge sum of intermediate data. Thus this paper analysis, how big data is used in diverse fields including health, agriculture, etc and focuses on issues and remedies for the particular concern by making use of distributed parallel processing with latest technologies.

Keywords: Hadoop Distributed File System [HDFS], Big Data in E-Health Service [BDeHS].

1. Introduction

The basic idea behind the phrase "Big Data" is that everything we do is increasingly having a digital trace(or data), which we (and others) can use and analyze. Simple activities like listening to music or reading a book are now generating data. Digital music players and eBook collects data on our activities. Your smart phone collects data on how you use it and web browser collects information on what we are searching for. Credit card companies collect data on where we shop and collect data on what we buy. It is impossible to imagine any task that does not generate data. Even many of our phone conversations are now digitally recorded including the conversations on social media sites like Facebook and twitter. We upload and share 100s of thousands of them on social media sites every second. We upload hundreds of hours of video images to YouTube and other sites every minute. In other aspects like, smart phone, it contains a global positioning sensor to track exactly where you are every second of the day, it includes an accelometer to track the speed and direction at which we are travelling at next level, we have smart TVs that are able to collect and process data, we have smart watches, smart fridges and smart alarms. The internet of things or internet of everything connects these devices so that e.g the traffic sensors on the road send data to your alarm clock which will wake you up have to leave earlier to make your meeting successful. Large data sets can be made from additional set of related data, as compared to separate smaller sets with same total amount of data, allowing correlations to be found to "spot business trends, then determine the quality of research, prevent diseases, link legal citations, compact crime, and determine real-time roadway traffic conditions". Therefore bigdata is data whose scale, diversity and complexity require new architecture, techniques, algorithms and analysis to manage and gain hidden knowledge and value from various resources. So big data is the understanding of realization of generate business intelligence by storing, processing the previously ignore data due to the limitations of traditional data management technologies. All together bigdata is new with massive collections of resources all put together in a data warehouse highly unstructured and used for social media and sentiment analysis with an aim to solve new older problems in a better way with different approaches, technologies, tolls and architectures. The best way to deal with varied data

sources and to meet objectives of the analytical process requires integration of varies new technologies

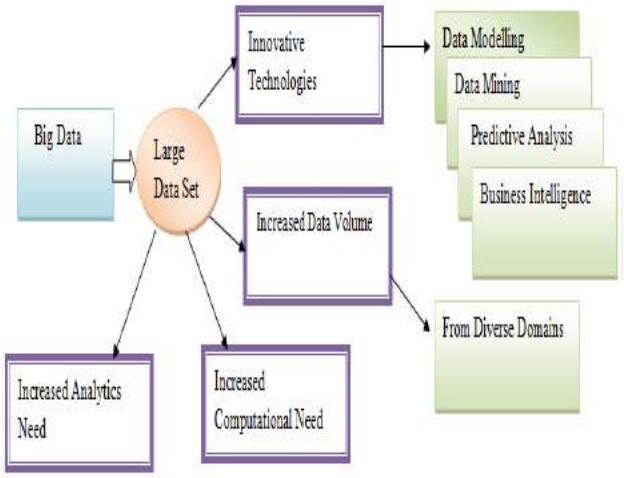

Figure 1. Big Data Technologies & It's Need

2. Big data characteristics

There are 5 characteristics of big data namely volume, variety, velocity, value and veracity.

Volume:- Denotes the capability to load and process the data e.g: from tera bytes to zetta bytes.

Volume = Length * Width * Depth

Variety :-Deals with different types of data and degree of structure that determine how to query the semi-structured data.

Velocity:- It is influenced by rate of data arrived in real time processing.

Veracity:-Veracity refers to messiness or trustworthiness of data

Value:- There is another V to take into account when looking at big data : Value -> Having access to big data is no good unless we turn it into values. Many industries are starting to generate amazing value from their big data. There is in need of large process required to form this big data in to value. The data might be activities, conversations, words, voice, social media, browser logs, photos, videos, sensors, etc…, therefore we need process like text analytics, sentiment analysis, face recognition , voice analytics, movement analytics, ecttoconvert this big data into value for this purpose, we need to analyze the structures, value, storage, source, quality and relationship of data to determine the value and appropriateness of dataset.

Table 1. Formation of Dataset

S.No	Structure	Storage	Source
1	Structured	Remotely accessed	Internal
2	Unstructured	Shared	External
3	Semistructured	Dedicated Platforms	Private
4	Table based	Portable	Public
5	Proprietary		

3. Big data challenges

Scalability for computations, Heterogeneity and incompleteness of data, scaling, timeliness(Query & indexing techniques to find suitable elements/records quickly), privacy, Human collaboration, claim to objectively and accuracy are misleading, and just because the data is accessible it doesn't make data to be ethical.

4. How is Big data actually used

Big data is used to better understand customers and their behaviors and preferences. Big data is also surprisingly used to optimize business processes, improvise personal quantification and performance optimization. The computing power of big data analytics enables us to find new cures and better understand and predict disease patterns. Thus big data plays prominent role in improving healthcare & public health. Big data uses video analytics, sensor technology and smart technology for improving sports performance along aspects on improving science and research, optimizing machine and device performance, improving security and law enforcement, improving and optimizing cities and countries and financial trading the five big data use cases focuses on big data exploration, operations analysis and data warehouse augmentation. There are eight major use cases of big data among eight use cases first one is optimize funnel conversion, then behavioral analytics, customer segmentation, predictive support, market basket analysis and pricing optimization, predict security threats, fraud detection and last one is industry specific. Big data analytics allows companies to track leads through the entire sales conversation process, from a click on an word ad to the final transaction, in order to

uncover insights on how the conversion process can be improved. Behavior, companies can learn what prompts a customer to stick around longer, as well as learn more about their customers characteristics and purchasing habits in order to improve marketing efforts and boost profits. In customer segmentation, by accessing data about the consumer from multiple sources, such as social media data and transaction history, companies can better segment and target their customers and start to make personalized offers to those customers. Next use case is predictive support through sensors and other machine-generated data, companies can identify when a malfunction is likely to occur. The company can then pre emptively order parts and makes repairs in order to avoid downtime and lost profits. Next use case is Market basket analysis and pricing optimization, by quickly pulling the data together from multiple sources, retailers can better optimize their product selection and pricing, as well as decide where to target ads. In predict security threats, big data analytics can track trends in security breaches and allow companies to proactively go after threats before they strike. Fraud detection concentrates more on financial firms that use big data to help them identify sophisticated fraud schemes by combining multiple points of data. In industry specific, virtually every industry has invested in big data to help solve specific challenges those industries face. Healthcare, for example, uses big data to improve patient outcomes, and agriculture uses data to boost crop yields.

Therefore big data exploration needs to find, visualize, understand all big data to improve decision making. Develop new business models with resulting increased market presence and revenue. Security/Intelligence extension enhances traditional security solutions by analyzing all type and source of under-leveraged data. Operation analysis analyze a variety of machine data for improved business results. This plays major role in decision making process and hence the benefits are gain real time visibility into operations, customer experience, transactions and behavior, Proactively plan to increase operational efficiency.

4.1 List of databases

List of databases that are active in storing bigdata are mysql DB, keyvalue DB, Document based DB, Column based DB, Graph based DB, RDF Graph data model, Dynamo, Voldemort, Bigtable, Cassandra, Hbase, Hypertable, Mongo database, SimpleDB, CouchDB and Dyrad

4.2 Data Types

Clinical data, claims, cost admission data, administrative data, research data, patient monitoring data, health data on web, pulseoximeter, blood pressure sensor, ECG, Wristband Kinnect, ICU Data, Images, Medical body area networks, sensing and transmittimg recorded measurements. Apart from this EEG, decomposing into wavelet form and GSR galvanic skin response. The medical image data's are computed tomography,magnetic resonance, imaging(MRI), Poison Emission Tomography (PET).

4.3 DataSources

The datas can be collected from internet, sensors, social media,bodysensor network and E-thin sensors. Apart from this data can also be collected from IOT, Crowdsourcing,social media, clinical dataset, hospitals,sensors, personal health records(PHR), EMR Electronic Medical records, ICU Data's, lab records, public health dataset and genetic dataset

4.4 Data Analytics

The data analysis process consists of predictive, prescriptive, diagnostic, descriptive, exploratoryand semantic analysis. The analysis process consists of preprocessing, cleaning,visualization and data preparation. The data preparation consists of transformation, aggregation, semantic integration and cleaning. Data preprocessing plays major operations like stemming, punctuation removal, stopward removal, integration, cleaning and redundancy elimination

4.5 Operations on data

Data capturing with the help of open libraries, data preprocessing (diarization techniques), data storage, data analysis, data visualization using R framework. The data visualized applications are alert detection, context extraction, repository management, pattern clusters, workflow processes, scorecards and interactive geospatial with 2D and 3D plots. With the help of above said operations many application in health sectors are analysed and they are listed below such as real time historical monitoring of patient's health, control of

infection, tracking and identification of patients, geofencing and vertical alarming.

But there are several issues when analyzing the data such as integrating,comparing, aaggregation, semantic content processing, data servicing, archival privacy restrictions, format complexity, interpreting, extracting or gaining knowledge.

Measures used are blood pressure, body temperature, heart rate, respiratory rate, blood glucose, blood oxygen, ECG signals, Height and weight measures, BGI, Sleeptime/Quality, meals(name, calories, compounds) and chest sounds.

Data management can be done by solving uncertainity, Query processing in realtime, constraints, Programming models, information extraction, information visualization, system architecture of machine learning and statistical methods. Successful applications in health sector by analyzing big data are summarized below like recommend therapy, Anamoly Detection,Prehypertension, Suggest diagnosis,Cardiovascular disease, blood pressure in pregnant, obesity and nutrition, aiding medical decision, type 2 diabetes, type 1 diabetes, prognotive computing (Active, Discrete, Atypical prognosis), cancer contagious, healthcare monitoring, neurological disease, parkinsons's disease, fraud detection, insurance health care, prediction of biomedical properties of nanoparticles, drowsiness/alertness detection, risk pregnancy care, ubiquitous operations, design of cancer chemotheraphy, indentify inpatient admission of individuals (primary diagnosis of cancer), identify severe asthma extracerbation, analysis and testing event logs(trace and flow analyses), prevents and controls the outbreak of diseases.

Data shedding techniques like pseudo-random sampling, compressive sampling and distributed source coding are done with optimization issues like stream processing and Batch processing with the online and outsourcing data's.

4.6 Medical image processing and techniques

Big data is the collection of large amount of data in the range of petabytes or terabytes of data collected from various stream of sources from different domains. One of the major domain of big data industry is health care. Health care industry is facing major challenge in handling digital data of numerous patients. This demonstrates the techniques used in processing the image data and also discuss the various applications of medical images.

4.7 Medical Data

ECG, EEG,ICU data, pulseoximeta, blood pressure sensor , galvanic skin response (GSR), Computed tomography(CT), magnetic resonance imaging (MRI), positron emission tomography (PET) and analyzing video.

4.8 Techniques used in processing Medical images

Case based reasoning, mapreduce, hdoop (pig), classification of images (using 5 classifiers like decision tree, naïve bayes, random forest and SVM), WEKA Waikato environment knowledge analysis, image miner, high performance computing, cloud data base, queue management, clustering of image data with 3 techniques like consensus, ensemble clustering. Along with this classification and correlation of images with Gaussian multiple instance learning techniques are applied . Image data analysis with hidden markov model and Gibbs model are done along with Machine Learning Techniques like SVM, PiSVM with STORM and Spark techniques. Support vector machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes comprises the input, making the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

4.9 Operations of Medical Imaging

Automatic registeration, segmentation, feature extraction, classification, image data management, query capabilities,Archiving(Image issues and organizing

image results), task assignment, data prefetching, overlapping of data/copy operations, analyses of components, algorithm for automatically prefetching and dividing images with fast training acquisition system. Construction of model for matching query to find nearest neighbor and classify label. Machine learning techniques like dimensionality reduction, recommendation system, malware clustering and image compression are major technologies and operations used in image classification.

4.10 Application of BIG DATA Images

High content analyses, face reconition in videos, remote sensing image processing,analyses of pathology images (queries using pipeline) ie., dataset of whole slide tissue is taken and content based image search and selection is done to analyse texture quantitative feature and manage mine results. Apart from this super resolution image reconstruction task, agricultural purpose (connecting image to wavelet tranformation method). This purpose consists of statistical, directional, periodical and low frequency statistical descriptors to texture information with wavelet transformation. The techniques used are image acquisition, classification task, image reconstruction task and mine selection task. For example X-ray of 14,410 images are processed for saliency template and image folding in online classification. This focuses on only fractural parts classifying only some portions of images, leftout remaining part, detect only salient features(regions) of images and reduce the effect of irrelevant regions focusing only on important part of picture alone(ROI Region of Interest) method, image compression method for compressing images based on hybrid PSO and GSA optimization algorithm.

Big data is a collection of massive and complex data sets and data volume that include huge quantities of data,data management capabilities,social media analytics and real-time data.Big data analytics is a process of examining large amount of data.In recent years the data mining applications become stale and obsolete over time. Incremental processing is a promising approach to refreshing mining results. It utilizes previously saved states to avoid the expense of re-computation from scratch. In this paper, we propose Energy Map Reduce Scheduling Algorithm, a novel incremental processing extension to Map Reduce, the most widely used framework for mining big data. Map reduce is a programming model for processing and generating large amount of data in parallel time. EMRSA is algorithm to provide more energy and less maps. Priority based scheduling is a task will allocate the schedules based on necessary and utilization of the Jobs. For reducing the maps, it will reduce the system work so, easily energy has improved. Hadoop Machine learning (ML) techniques have shown impressive performance in solving real life classification problems in many different areas.Models such as Naïve Bayes and Support Vector Machine (SVM) techniques are used. Here we use SVM classifier, support vector machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. Naïve Bayes is a conditional probability model which can be extremely fast relative to other classification algorithms.Hadoop is an open source framework that allows to store and process big data in a distributed environment across clusters of computer using simple programming knowledge, It is designed to scale up from single servers to thousands of machines each offering local computation and storage. Finally, results shows the experimental comparison of both the algorithms involved.

According to Jeffrey Dean &et.al[20],MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. Our implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on

thousands of machines. Programmers find the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google's clusters every day.

According to Matei &et.al[21] Zaharia Mosharaf Chowdhury Tathagata Das**,**We present Resilient Distributed Datasets (RDDs), a distributed memory abstraction that allows programmers to perform in-memory computations on large clusters while retaining the fault tolerance of data flow models like MapReduce. RDDs are motivated by two types of applications that current data flow systems handle inefficiently: iterative algorithms, which are common in graph applications and machine learning, and interactive data mining tools. In both cases, keeping data in memory can improve performance by an order of magnitude. To achieve fault tolerance efficiently, RDDs provide a highly restricted form of shared memory: they are read-only datasets that can only be constructed through bulk operations on other RDDs. However, we show that RDDs are expressive enough to capture a wide class of computations, including MapReduce and specialized programming models for iterative jobs such as Pregel. Our implementation of RDDs can outperform Hadoop by 20× for iterative jobs and can be used interactively to search a 1 TB dataset with latencies of 5–7 seconds.

According to Russell Power Jinyang Li,Pi &et.al[22] is a new data-centric programming model for writing parallel in-memory applications in data centers. Unlike existing data-flow models, Piccolo allows computation running on different machines to share distributed, mutable state via a key-value table interface. Piccolo enables efficient application implementations. In particular, applications can specify locality policies to exploit the locality of shared state access and Piccolo's run-time automatically resolves write-write conflicts using userdefined accumulation functions. Using Piccolo, we have implemented applications for several problem domains, including the PageRank algorithm, k-means clustering and a distributed crawler. Experiments using 100 Amazon EC2 instances and a 12 machine cluster show Piccolo to be faster than existing data flow models for many problems, while providing similar fault-tolerance guarantees and a convenient programming interface.

According to Grzegorz Malewicz &et.al[23],Many practical computing problems concern large graphs. Standard examples include the Web graph and various social networks. The scale of these graphs—in some cases billions of vertices, trillions of edges—poses challenges to their efficient processing. In this paper we present a computational model suitable for this task. Programs are expressed as a sequence of iterations, in each of which a vertex can receive messages sent in the previous iteration, send messages to other vertices, and modify its own state and that of its outgoing edges or mutate graph topology. This vertex centric approach is flexible enough to express a broad set of algorithms. The model has been designed for efficient, scalable and fault-tolerant implementation on clusters of thousands of commodity computers, and its implied synchronicity makes reasoning about programs easier. Distribution related details are hidden behind an abstract API. The result is a framework for processing large graphs that is expressive and easy to program

According to Svilen R. Mihaylov Zachary G &et.al[24],In today's Web and social network environments, query workloads include adhoc and OLAP queries, as well as iterative algorithms that analyze data relationships (e.g., link analysis, clustering, learning). Modern DBMSs support ad hoc and OLAP queries, but most are not robust enough to scale to large clusters. Conversely, "cloud" platforms like MapReduce execute chains of batch tasks across clusters in a fault tolerant way, but have too much overhead to support ad hoc queries. Moreover, both classes of platform incur significant overhead in executing iterative data analysis algorithms. Most such iterative algorithms repeatedly refine portions of their answers, until some convergence criterion is reached. However, general cloud platforms typically must reprocess all data in each step. DBMSs that support recursive SQL are more efficient in that they propagate only the changes in each step — but they still accumulate each iteration's state, even if it is no longer useful. User-defined functions are also typically harder to write for DBMSs than for cloud platforms. We seek to unify the strengths of both styles of platforms, with a focus on supporting iterative computations in which changes, in the form of deltas, are propagated from iteration to iteration, and state is efficiently updated in an extensible way. We present a programming model oriented around deltas, describe how we execute and optimize such programs in our

REX runtime system, and validate that our platform also handles failures gracefully. We experimentally validate our techniques, and show speedups over the competing methods ranging from 2.5 to nearly 100 time

5. Impacts of Big Data Analysis

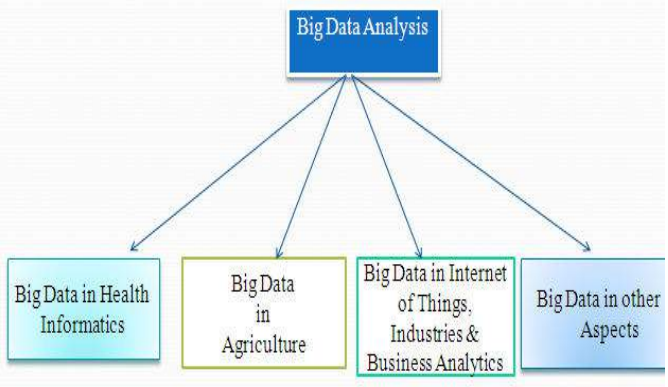

Figure 2. Big Data Analysis

Before going into several domains, let us discuss how big data and its technologies are helping real time movement of data. Ishwarappa & Et.al [1], main concepts of this paper focus on introduction about big data, technologies used and its challenges. For example, Big data is a collection of huge and compound data sets. Immense quantity of data management and organization capabilities are done. IT is a collection of high volume of digital mixed data. It contains open source disseminated information dispensation with HDFS [Hadoop Distributed File System]. HDFS is defined as inter relative & correlated free cohesive platform. The main challenge that discussed in [1], is capturing, analyzing, storing and consolidating the available data to make it useful or decision making and improved business intelligence process. The technologies used [1], are discussed in summarization table given below. Srivathsan & Et.al [2], demonstrate how prognotive computing is related to big data analytics. Prognotive computing is a process of accumulating dealing and examining extremely large amount of structured and unstructured clinical information stemming out from a wide range of experiments, summary collected from hospitals and laboratories including pharmaceutical companies or even from social media. This paper [2], discusses why big data is used in prognosis. Prognotive computing depends on the availability of information from different sources/or domains hence commitment among all organization is a valid requirement for setting up big database or warehouse. This paper [2] uses patterns and pattern recognition techniques to collect information about disease and groups them into cluster and manages each cluster with tree structures and group. They have divided the entire process into 3 steps. Active scanning, discrete prognosis and atypical prognosis. First phase active scanning collects details about disease from all the e-health services in online real time aspect and suggest the patients to approach best e-case professionals based on their disease. Second phase discrete prognosis, takes the patient data as input and provide medications for curing third stage called atypical prognosis forecasts future conditions of patient body and provide him some suggesting to avoid future diseases. The very big changes to face in Roger Hernandez & Et.al [3], is the commitment among all domains for the setup of real time big data warehouse and it is difficult to implement prognosis tor healthcare process which includes attention lot of research issues to be taken into account.

5.1 Data Modelling Technique

Roger Hernandez & Et-al[3] propose query driven data modeling technique which is a transparent data management technique that replicates data's into different model sand this QDDM provides greater response by matching the queries with well-performing model option. QDDM models scheme automatically and removes inconvenience and short coming user for maintaining in the data consistency. QDDM choose the best appropriate model or layout for managing heterogeneous replication mechanisms. The method involves receiving the incoming request in the form available model to serve query with improved performance. The main drawback in this model is that cases high overload. Florian Endel and Harald[4] proposes a new concept called data warehousing which helps in making data useful again. There is a need to analysis data in every day's business needs. Data wrangling is not only about renovating and clean up procedures but also includes other aspects like data quality, integration of different sources from different domains , replication process for handling data provenance. All these above specified aspects have to be considered. This applies realistic rational routines of interactive and multidisciplinary process. The main objective of data wrangling is to enable deep

understanding of the content, structure, quality issues and necessary transactions as well as suitable tools/technological resources are needed. The main motivation of this paper is to collect data from more/less structured sources and preparing it foo virtualization, modeling or permanent storage is a necessity of this proposed work. The main challenges of this work are ensuring that appropriate data with proper size and encoding format is collected and several dimensions of data quality with divergent structures, coding conventions are linked together. Altogether handling data uncertainty, uncleaned data, fault & error tolerance with transformation/ editing also needs to be considered.

5.2 Big Data in Cloud

Victor Fernandez & Et al [5] manages big data with distributed infrastructure with remote access control access point [DIRAC]. The main aim of this paper is to access multiple resources from divergent fields, managing jobs and balancing the work load by partitioning the task into several components, monitors them and updates the result. This proposed work, as name implies provides access not only to grid or cloud resources but also provides access to big data resources from the same DIRAC environment. Load balancing technique used here is to aggregate distributed resources in an efficient manner. Therefore DIRAC is a platform providing a complete solution to scientific communities accessing distributed resources including clouds, grids and local clusters. This enables to access data scattered in different divergent geographical areas such as access to grid resources, including a centralized data catalog to integrate & combine big data clusters in a coherent way. The 2 main components of DIRAC workload management is intensification and diversification. Intensification focuses in local region to ensure convergence to optimal and Diversification focusses global region along with dynamic constraints. As a result, the proposed work provides single scheduling mechanism for jobs with very different profiles. To achieve overall optimization it organizes remaining rending jobs in task queue with job priority assigned to each queue with similar requirements. As a future work , this paper concentrates on workload management and scheduling because this enforces each cluster with only waiting job. Dilpred singh & et.al [6] discussed about the platform for bigdata analytics. This paper mainly focuses on different platforms that are used for bigdata analytics. In depth-analysis of different platforms are done based on several metrics. Two types of platforms say hardware and software platforms with various platforms like scalability, data I/O rate, fault tolerance, realtime processing and data size are considered. Issues in application level and system level requirements are sorted out. This paper implements K means clustering algorithm on various platforms are discussed. At application/algorithm level, the issues like how quickly do we get the results, how big is the data to be processed, does the model requires several iterations or single iterations are sorted out and at system/platform level there are several issues that should be taken interest such as, is the rate of data transfer critical for this application, is there is a need for handling hardware failures within the application, whether we need more data processing capability in future ,etc are considered. Apart from this various issues like scalability, data I/O performance, etc should be handled. Horizontal vs Vertical scaling must be done with high input/output rate of huge real time data with fault tolerance techniques, providing support of iterative tasks in real time processing. The main drawback of this paper is that there is a need for investigating more algorithms in both hardware and software platforms. Future work leads to the possibility of combining multiple platforms to solve application problems. Nagwani [7] demonstrates how to summarize large text collection using topic modelling and clustering based on mapreduce framework. This paper focuses on document summarization process. This leads to better understanding of text documents. But this is challenging and time consuming process. Summarizing single text document is easy then summarizing multiple text document. This requires semantic similarity based clustering and Latent Dirichlet Allocation mechanism.

The entire task is completed in four stages , First stage is document clusteing stage where similar texts are clustered in multiple documents and make it ready for summarization. Second stage is topic modelling technique(LDA) , where each individual text document generates cluster topics and realated terms belonging to those cluster topics. At third stage global frequent terms are collected and finally duplicate sentences are identified and removed. The main aim of this paper is to maintain high compression and retention ratios. Nemanja & Et.al[8] compares data flow and control flow models. This is done by taking both quality and quantity aspects into account. They proposes a new

venture for benchmarking which takes execution time (speed ups), power reductions and space savings with considerable results.

5.3 Big Data in E-Health

Matthew Herland & Et.al [9] takes real world of medical data from all levels of human existence to advanced health care and medical practices. This paper combines information science and computer science to make research in health informarics. The process range from data acquisition , information retrieval, storage and analytics. Health informatics employs data mining techniques because it contains data collected from different domains such as Bio Informatics, Image Informatics, Clinical Informatics, Public Informatics and transactional Bio Informatics. Computational power becomes major drawback. More efficient and accurate methods needs to be developed and Health Informatics becomes Really Big Data (RBD). Hence testing is required before processing to next level. Maryam M Naja & Et.al [10] illustrates how deep learning applications make use of big data and how its challenges and analytics can be done. This paper collecets massive amount of domain specific information. This proposed work indicates how specific areas of deep learning can be improved to reflect certain drawback in big data analytics. The main problem is data abstraction, Information retrieval, High dimensional data and distributed computing. Vishal & Et.al [11] showcases how skewness can be handled in big data using map reduce. Skewness refers to uneven partioning of reducers ie., some reducers have more data to process than other reducers. Skewness occurs in mappers, reducers and strangler nodes. Skewness in mappers occurs in such a way that mappers will take same size data and it will split the input data's into partitions. But the complexity for those partitions may vary. The output of mappers will be taken as input to the reducers. Very large key size is required to process before ordering the partitions. Simillarly some nodes have less compute powers than other nodes hence called straggler nodes. This causes skewness and hence reduces the overall performance. This paper mainly concentrates on efficient load balancing and partition techniques to remove skew and improve the performance and increase the number of jobs done. The main drawback here is the difficulty to mitigate skewness in copy and sort phases. Benjamin Gufler& et.al [12] focuses on key challenges such as how to process massive amounts of data. Key contributions of this paper is to develop the cost model to distribute the workload. This consists of 2 process called fine partitioning and dynamic fragmentation. Fine partitioning splits the information into fixed number of partitions and estimates the cost of all partitions and tries to balance the clusters such that all the reducers will roughly require same time for processing the information. The cost model is designed in such a way that no reducer remains idle and becomes well balanced with minimum execution time to complete the job. Skewness occurs in keyed frequency, tuple size and execution time. Some key appears more frequently in intermediate data . Some application holds large complex objects and single large clusters requires more execution time then many small clusters. Dr.W.Liu & Et.al [13] demonstrates how big data is used in E-Health Service. Big data health service [BDeHS] collects resources from electronic medical records, Mobile Health Records and Patient Health record. Hence we are in need of STORM, free open source distributed real time technology. This helps for making details about diagnosis, Procedures, Medications, Digital images, medicine supplies, lab results and billing. The streams of data integrates with queuing and database technology. It will make process in arbitrary ways and guarantee the data consistency. It is scalable and fault tolerant system that takes into account parallel processing, data segmentation, summarization and policy enforcement. Gajanayake.R & Et.al [14] demonstrates secured E-Health data retrieval in DaaS. Their proposed work is simulated in three types of environments and makes use of cloud services. This makes use of bloom filter to remove the unwanted or unclear messages. This takes into account client response time, processing time and database processing time. Herland .M & et.al [15] surveys clinical data mining applications on big data in health informatics. The main objective is to improve the length and quality of life for patients with data mining and machine learning applications. Health informatics is required to answer various questions. The main advantage is earlier prediction of disease and advanced dynamic protection.

5.4 Big Data in Agriculture

Ludena & Et.al [16] Demonstrates how big data approach is used in ICT agriculture approach. This paper mainly focuses on issues like how to collect agricultural data's, how to maintain security and provide privacy

protection to farmer's individual data's, relationship among them, result interpretation, dataset equivalence and cleaning of uninterested data's. This paper mainly focuses on metadata generation. The key contributions are creating a system that provides healthy food suggestions to the user; provide food requirements of daily needs. There is necessity to summarize data set because it contains information such as weather change, price instability, ethical behavior of farmers, etc. The main difficulty is to construct the data acquisition system. Hirafuii & Et.al [17-19] talks about strategy to create big data and how to make use of distributed processing and cloud computing in agricultural decision making system. The main aspect is that since data's are created naturally, farmers are in need of application to assist them in making decisions when environment condition and climate changes. Multiple equipments are needed like field sensors for field monitoring, network cameras, routers, solar cells and sensors. Maintenance becomes a major drawback. So the work has been configured in Hadoop map-reduce framework where the job execution is done by configuring slave nodes, loading the natural data's into HDFS, executing the job with reducers and final report is made. Finally results are transferred to the web portal. This helps not only in decision making but also for increasing the production. Future work of this paper aims in reducing the configuration time. So overall process requires steps like gathering the information, sharing all transaction information, perform specific big data analysis based on mobile data and data which is at rest. It will interpret the results and improve the accuracy and effectiveness of agricultural system.

Table 2. General Issues

Outcome of General Analysis			
Big Data in Health Informatics	Big Data in Agriculture	Big Data in Internet of Things and Business Analytics	Big Data in other Aspects
Limitations/Drawbacks			
• Security • Data Dissemination • Data Integration	• Analysis strategy. • Privacy and data protection. • Data Cleaning	• High Overhead • Dirty Data • Data Uncertainty • Error Tolerance • Transformation and Editing	• Low retention ratio • Network traffic • Information Retrieval

6. Conclusion

There is no single definition to define the full capabilities of big data. It is the term for collecting & organizing very large data sets that is so large and difficult and becomes complex to process using on-hand database management tools or traditional data processing tools. The common issues in big data include capture, curation, storage, search, sharing, transfer, analysis and visualization. This paper mainly focuses on how to mitigate the data skew problem in big data. Future work tends to reduce the data skew problem to greater extent in various fields particularly in medical and agricultural domain with the aid of technologies used in Hadoop distributed file systems

7. References

1. Ishwarappa, Anuradha, A Brief Introduction on Bigdata 5Vs Characteritics and Hadoop Technology, Elsevier, Science Direct, Volume 48 , (2015) 319-324
2. Srivathsan M , Yogesh Arjun K, Health Monitoring System by Prognotive Computing using Bigdata Analytics, Elsevier, Science Direct, Procedia Computer Science ,Volume 50 ,(2015) 602- 609.
3. Roger Hernandez , Yolanda, Eduard and et.al, Roger Hernandez , Yolanda, Eduard and et.al, Automatic Query Driven Data Modelling in Cassandra, Elsevier, Science Direct, Procedia Computer Science.
4. Florian Endel, Harald Piringer, Data Wrangling:Making Data useful Again, Elsevier, Science Direct, IFAC –PapersOnLine, Volume 48-1 (2015) 111-112.
5. Victor Fernandez, Victor Mendez, Tomas F.Pena,Federated BigData for Resourse Aggregation and Load Balancing with DIRAC, Elsevier, Procedia Computer Science, Volume 51, 2015 pages, 2769-2773
6. Dilpreet Singh and Chandan K Reddy, A Survey on Platforms for Big Data Analytics, Journal of Big Data, Springer Open Journal, 2014 , Volume 1:8.
7. N.K Nagwani, Summarizing Large Text Collection using Topic Modeling and Clustering based on Map Reduce Framework , Journal of Big Data, Springer Open Journal, 2013, Volume 2:6.
8. Nemanja, Veljko, Jakob and Anton Kos,Paradigm Shift in Big Data Super Computing: Data Flow Vs Control Flow, Journal of Big Data, Springer Open Journal, 2015, Volume 2:4

9. Matthew Herland, Taghi M, and Randall Wald, A Review of Data Mining using Big Data in Health Informatics, Journal of Big Data, Springer Open Journal, 2014, Volume 1:2
10. Maryam M Naja, Flavio, Taghi, Naeem Seliya & et.al, Deep Learning Applications and Challenges in Big Data Analytics, Journal of Big Data, Springer Open Journal 2015, Volume 2:1
11. Vishal A.Nawale, Priya Deshpande, Skewness a challenge for Map reduce in Big Data,AVCOE,Sangamner,IPGCON-2015
12. Benjamin Gufler, Nikolaus Augsten, Angelika Reiser and Alfons Kemper, Handling Data Skew in Map Reduce,CLOSER 2011 - International Conference on Cloud Computing and Services Science
13. Dr.W.Liu, Dr.E.K.Park,Big Data as an E-Health Service ,IEEE, Computing, Networking and Communications, Feb 2014, Pg. No-(982-988)
14. Gajanayake.R, Sahama.T, Secured E-Health data retrieval in Daas and Bigdata, IEEE, Networking, Services and Applications(Healthcom),Dec 2013
15. Herland .M, Boca Raton, Survey of Clinical Data Mining Applications on Big data in Health informatics, IEEE, International journal of Machine Learning Applications DEC 2013, Pg.No(465-472)
16. Ludena, R.D.A, Ahrary,Big data Approach in an ICT Agriculture Project, IEEE, Awareness Science and Technology and Ubi-Media Computing, NOV 2013,PG NO:261-265
17. Hirafuii.M, A Strategy to Create Agriculture Big Data , IEEE, Global Conference (SRII), 2014 Annual SRII, April 2014, Page(s):249 – 250
18. Akio Goya, Risse de Andrade, M. ; Carvalho Zucchi, A ; Mimura Gonzalez, N, The use of Distributed Processing and Cloud Computing in Agricultural Decision Making Support Systems , IEEE ,Cloud Computing (CLOUD), June 27 2014-July 2 2014 , Page(s):721 – 728
19. Ahrary, A. Ludena, R.D.A,Big Data Approach to a Novel Nutrition-Based Vegetable Production and Distribution System, IEEE, Computational Intelligence and Cybernetics (CYBERNETICSCOM), Dec 2013,Pg.No -131-135.
20. Jeffrey Dean and Sanjay Ghemawat, "Mapreduce: simplified data processing on large clusters," in Proc. of OSDI '04, 2004.
21. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in Proc. of NSDI '12, 2012.
22. R. Power and J. Li, "Piccolo: Building fast, distributed programs with partitioned tables," in Proc. of OSDI'10, 2010, pp. 1–14.
23. G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: a system for large-scale graph processing," in Proc. of SIGMOD '10, 2010.
24. S. R. Mihaylov, Z. G. Ives, and S. Guha, "Rex: recursive, deltabased data-centric computation," PVLDB, vol. 5, no. 11, pp. 1280–1291, 2012.
25. Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein, "Distributed graphlab: a framework for machine learning and data mining in the cloud," PVLDB, vol. 5, no. 8, pp. 716–727, 2012.
26. S. Ewen, K. Tzoumas, M. Kaufmann, and V. Markl, "Spinning fast iterative data flows," PVLDB, vol. 5, no. 11, pp. 1268–1279, 2012.
27. Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst, "Haloop: efficient iterative data processing on large clusters," PVLDB, vol. 3, no. 1-2, pp. 285–296, 2010.
28. J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox, "Twister: a runtime for iterative mapreduce," in Proc. of MAPREDUCE '10, 2010.
29. Y. Zhang, Q. Gao, L. Gao, and C. Wang, "imapreduce: A distributed computing framework for iterative computation," J. Grid Comput., vol. 10, no. 1, 2012.
30. S.BrinandL.Page,"Theanatomyofalarge-scalehypertextual web search engine," Comput. Netw. ISDN Syst., vol. 30, no. 1-7, pp. 107–117, Apr. 1998.
31. D. Peng and F. Dabek, "Large-scale incremental processing using distributed transactions and notifications," in Proc. of OSDI '10, 2010, pp. 1–15.
32. D. Logothetis, C. Olston, B. Reed, K. C. Webb, and K. Yocum, "Stateful bulk processing for incremental analytics," in Proc. of SOCC '10, 2010.