Performance Evaluation Of Application Specific Big Data Systems With Multi-domain Big Data Framework Using Machine Learning

Dar Masroof Amin¹, Dr Munishwar Rai²

¹²M. M. Institute of Computer Technology & Business Management, Maharishi Markandeswar (Deemed to be University), Mullana, Haryana 133203, India
¹masroofiju13@gmail.com, ²munishwar.rai@mmumullana.org

Abstract. The world is getting adopted towards Big Data analytics and related tools which has increased demand for secure, reliable and efficient technology for processing large amount of data flowing on internet. The analytics of datasets has provided opportunities in industries and academics to look into the drifts while creating these large datasets. The developments in network technology and other data analytics systems has increased the data collection in various domains i.e. social networking, medical care, agriculture and horticulture, business and finance, educational industries and smart cities. The use of such amount of data through process of analytics using advanced techniques generally helps in predicting the trends in future and reviewing the change with optimal precision. The diverse nature of data, quantity and speed of creation is very much high with obsolete security procedures which make it un-reliable to use. These provide scalable and efficient results as compared to the traditional platforms and techniques. The research already done focusses on independent domain specific ICT applications. But less work has been done in making a collective approach to deal with growing use of data, systematic storage, extraction of domain related datasets, security and analytics thereof. The research presented evaluates the big data platforms with novel multi-domain big data framework using the available techniques for security and other challenges.

Keywords: Big Data, Internet of Things(IOT), Artificial Intelligence (AI), Machine Learning (ML), Smart Data, Hadoop Distributed File System(HDFS), MapReduce.

1. Introduction
The Big data is outcome of growing digital technology and related applications. The addition of technology in the form of high definition computers mobiles and sensor networks is creeping with speed into the lives. The realtime inter and intra communication among machines and humans is creating enormous amount of data with high veracity, velocity and variety [8][10]. These technological advancements have created exponential increase in use and storage of large amount of data [1][2]. The data created is having variety, heterogeneity and speed that needs to be processed efficiently for useful mining of valuable inferences. Therefore, development of novel tools and techniques has become important to explore the bottlenecks while creation of massive amount of data and analytics thereof. The big data is created in time series manner over a period of time with visible foot prints to make future decisions [3]. And same information is used for establishing the relationship between various attributes, extraction of patterns and development of future trends.
The engines that process this huge amount of data have provided an opportunity to make improvement in performances through reliable and optimal movement of data while executing the applications. The scheduling of type locality-aware introduced by researchers provides a network transfer with low cost and provides scheduling of jobs to the nodes which are nearest to the stored data [4]. The in-memory process of computing further enhanced the optimal decrease in movement of data and whenever in need, ensured intra-tasks confinement of data in memory [5]. The cost of movement of data has been reduced, timely evaluation, the computations are invoked only when needs, was all handled by the Big Data technologies for optimization process viz regrouping, computing at the time of necessity. Frameworks such as Spark and MapReduce [6] have taken place as primary tools for big data analytic and many more are emerging into the race like Dask [7].

Because of increase in volumes and heterogeneity of data, the most of the scientific areas like biotechnology, physics, astronomy and bioinformatics have entered into the big data domain. The algorithmic techniques are encompassed in machine learning which makes the systems to work with problems without following any programming language concept. The research presented evaluates use of AI, machine learning tools and analytical models in a scalable multi domain framework. The research presented evaluates the framework proposed in [9] using existing techniques proposed from time to time. The analytics solutions provided in various research articles addresses the problems in a specific domain like health, traffic etc. Similarly, there are separate solutions for application interoperability in which the data in domain applications become source for analytics [10]. However, the research presented in [9, 11] provides a comprehensive solution for taking out the project in universal domain by adhering to the rules of artificial intelligence and machine learning.

The rest of the organization of the research presented as: The introduction provides the details about subject under consideration and motivation for taking this research. The section-2 provides the related work studied in context of the research presented. The general concept about integrated smart cities project with multiple domains is discussed in Section-3. The Section-4 provides the general outlook of existing big data systems. The evaluation analysis regarding framework proposed [9] with the traditional techniques and prime contribution of this research are given in Section-5. Finally, the conclusion and future scope is discussed in Section-6.

2. Related Work

The researchers presented evaluates Hadoop, Flink and Spark using the Big Data input workloads by taking consideration the metrics like scalability and performance. In addition, the parameters have been modified for looking into the behavior of the frameworks based on HDFS block size. A comparative study has been done by considering the parameters associated with scalability and performance of Hadoop, Spark and Flink on different Big Data workloads. The frameworks behavior has been characterized through modification of parameters viz block size, input, thread configuration and interconnect network. The research shows overall reduction in execution times on average basis [29].

In order to quantify performance bottlenecks authors developed a blocked time analysis for distributed computation frameworks. The performance has been studied on three workloads and it was found that jobs on CPU are bottlenecked and performance of network has little impact on completion time of a job. The research studied analytical frameworks for performances on smaller workloads. The research provided insight to develop methods to work on more performance improvements [37].

The use of Hadoop for data analysis with systems having varied capacities has been studied. The research creates data through algorithms and analysis of data is done on multiple systems for understanding the underlying constraints and limitations. The study evaluates that using difference processors and memories can help in improving the system performance and also reiterates that Hadoop has limited applicability on single memory systems [38].

The research presented describes how spark is more efficient than Hadoop and also provides the general overview of both. The spark performs in memory computations, stream processing and interactive querying and is stronger than Hadoop. The research also provides details about how Hadoop works on
distributed storage environment though spark is hundred times faster. The spark provides advanced analytics applications and is used on top of Hadoop for distributed file system [39].

The authors provide a comparative study of various machine learning algorithms. The sole purpose is to look for optimal classification of data so that there will be proper analytics for treating a disease. The research obtained results using real analysis and simulation showed that Random Forests, SVM (Support Vector Machines) and Naïve Bayes provides better results that KNN (k-Nearest Neighbour) and LDA (Linear Discriminant Analysis) for small as well as large samples [40].

The authors presented overview of big data analytics life cycle in manufacturing IoT (Internet of Things). The research summarized that there are three phases in the general life cycle viz data acquisition, pre-processing, storage and analytics [41].

The research presented a methodology to provide evaluation of data driven performance models in energy sector. There is interpretable learning methodology is developed. A new metric trust is proposed to provide alternative to traditional metric for evaluating performance. The results have been obtained by implementing methodology on actual operational data. The evaluation shows that more intelligent management systems can be developed with user-friendly environment. The Gower’s dissimilarity coefficients and clustering analysis has been integrated to enhance the quality of data analysis [42].

The authors presented forecasting using ensemble models in big data time series. The random forests, decision tree and gradient boosted trees’ ensemble has been proposed for providing better results than previous applications big data. The problem is divided in sub-forecasting problems. These sub-problems are solved using machine learning methods using big data Apache Spark engine. The results have been performed on static as well as dynamic ensembles and have performed well in both [43].

The research aimed to provide an overview regarding how big data analytics has improved systematically the process of decision making. The general techniques used by data scientists for analytics purpose for meaning results are through modeling, data mining, visualization, machine learning and statistical analysis [44]. The various stand-alone application domains are using machine learning for process of predictive analytics [45]. The machine learning is a complex field in which various frameworks, tools and techniques can be used to address the growing use of data and fusion [46].

3. Sample Smart Cities Project.
The smart cities project architecture presented in [10] provides the basic structure framework to simulate the activities in smart city. The architecture presented is divided into three modules viz platform layer, data processing layer and security layer. The project is considering concept of city smartening by taking into consideration the growing urbanization, where ICT plays a major role in mitigating the various activities in different domains. The ICT plays an important role as the top organizations has gained interest in using big data analytics in extracting the hidden patterns for making future decisions. The above architecture selects data from various sources and pre-processes the same for further analytics. The design changes as proposed in [9] and further utilization of machine learning methods as proposed in [17] is going to develop a big data system that will revolutionize the smart data analytical project. The usability of AI tools will have major effect on human decisions and also address the legal and general ethical inferences. The addition of system specific machine learning algorithms in future is goings to provide high definition real time results. Figure 1 provides the general architecture of smart city project.
The above architecture selects data from various sources and pre-processes the same for further analytics. The design changes as proposed in [9] and further utilization of machine learning methods as proposed in [17] is going to develop a big data system that will revolutionize the smart data analytical project. The usability of AI tools will have major effect on human decisions and also address the legal and general ethical inferences. The addition of system specific machine learning algorithms in future is going to provide high definition real time results.

The spontaneous increase in number devices connected to the interned has made data storm on internet and use of ICT for making the various daily life activities has gave birth to the IoT. This led to creation of big data and its widespread use provided real time solution to the problems. The analytics of these large number of data clusters has provided opportunity to creep deep into the problems and issues. The kind of data usage provide insight to develop AI models for future predictions by using archival data correlations. In addition, the linkage has been created between the realtime data and archival research data for inline solution to the problem arising out in a particular domain. Moreover, combination of various algorithms dealing with neural networks has led to the development powerful models. These flexible and scalable models provide the base for deep learning concept which are trained iteratively for solving complex problems. AI models also provides services through centralization and data source pooling. The AI models does it by aggregating the heterogeneous data and analyze it through machine learning tools. The development of these tools and its integration with smart data analytics models will transform into an effective advancement. And for non-standardized data, methods are available to adjust the irregularity and bias [12].

The big data is efficiently handled by deep learning through building flexible relationship among complex variables in trainable manner [13]. The deep learning through multiple sequential layers facilitates high dimensional analytics of data [14]. With technological advancements, deep learning is different from other machine learning concepts and has capacity to handle large number of variables.

To get the general analytical solutions out of big data, there is need proper labeling and curation. The data generated is of diverse nature with heterogenic properties. The labels should be in a format to specify the domain for performing machine learning. The invalid labelling will have an impact on various metrics viz accuracy, efficiency etc. that is supposed to get attained by machine learning process. Therefore, before building any model the proper categorizing of data to their respective domains for ensuring reliability [9][15]. The data curation at node level inside a domain will help in reclassifying into logically sub-modules which helps in accurate predication while making data pass through machine learning algorithm. In order to provide logical conceptualization in machine learning, the proper semantics and format needs to be developed for data in nodes [11][16]. The high definition neural
networks ignore the concept of idealism of datasets which is making it provide machine learning even for income data points.

4. Existing Systems.

The existing Big data frameworks are classified on basis of processing which can be of type streaming, batch and hybrid. The tools incorporated with frameworks provide diverse functions and services which include monitoring of workflow, storage management, and optimal resource utilization file system. These tools are generally categorized as analytical, management and storage modules [47].

The basic technologies that are currently dealing with problems associated with big data and are the prime technical sources for performing analytics on large datasets. The processing of heterogeneous data for purpose of extracting future trends from various big data sources viz sensors, social networking etc. These technologies are Data Center, Internet of Things (IOT), Cloud Computing, and Hadoop [20]. These technologies plays important role in dealing with big data related systems.

The data center platform provides solution for concentrated data storages which provides services related to acquisition, organizing, management of data and corresponding functions. The centralized data centers are obsolete so far as growth of real time heterogeneous streaming of data is concerned. The infrastructure needs the timely updation for techniques and technologies for big data [21].

The Internet of Things is the basic framework for set of objects that are solely representing and are recognized as independent devices on internet. These independent devices are digital camera, sensors, smartphones, tables, palmtab etc. [20]. The interconnection of these devices on web provides smart services and processes with regard to our day to day life activities. These smart devices on the internet produces big data and delivers multiple facilities, which produces massive amounts of data and information [49]. The sensor networks growth and use of machine learning is making the realtime systems to produce huge set of data in smart cities, houses and medical area [50]. The data created by IT devices and IoT devices is unstructured, or semi structured and growth of creation of this data will reach to one trillion by 3030 [50]. The growth of technology is lagging than the speed of creation of data. The current designed systems are obsolete so far as scalability and availability of systems are concerned.

The Apache Software developed Hadoop technology, an open source project for batch processing. The sub modules of the project are highly efficient for distributed computing. The projects commonly look into the big data projects viz social network click streams, filtering of spam’s, network surfing, intrusion detection etc. The Hadoop works with HDFS(Hadoop Distributed File System) file system and corresponding programming module MapReduce. The HDFS stores data in clusters in a distributed fashion. The HDFS is fault tolerant, scalable, portable distributed storage system written using Java language. The nodes in HDFS are of two type, the master name node and data nodes. The name node is used to keep meta data, file system tree and namespace [22]. The unstructured and semi- structured data is used as input. The programming paradigm has been introduced by Google namely MapReduce which is divided into two sub modules Map and Reduce and performs parallel processing of data distributed on various clusters. The modules main activity is generating large datasets and their processing. The Map sub module divides master node data into smaller job modules. The smaller problems are processed by a worker node through Job Tracker and results are stored in local file system. Accordingly, reducers access the data from local file system for further processing [23]. The (key, value) pair format is the output of Mapper and is a sorted list. The list is shuffled for next phase and latter Reducer module integrates the input data for analysis. The presence of multiple reduce makes the process of parallelization for aggregation and Job Tracker controls the tasks which are executed on worker nodes [24]. The Figure.2 provides the big data basic methodology which provides the prime models to be followed while developing a system that incorporates Big Data and Computational Intelligence [25].

The rapid development in applications and social media systems is creating tremendous amount of data in a minute, which becomes an important source in research areas. In order to utilize the data, there is
need of realtime analytical solutions in the form of algorithms and technology. The methods discussed above are data driven technological solutions for optimal use of big data. There are few issues that are currently arising out of the technologies discussed in previous section. The issues mainly pertain to data acquisition, storage, parallel processing and finally the analytics to get meaning patterns to deal with realtime problems and future predictions. Moreover, there is need to exploit the new methods and techniques to provide smart solutions for enhancing the quality of services and standards. In addition, existing frameworks and soft solutions generally deal specific domains and applications. The proposed system in [9][11] which provides a detailed framework to deal with privacy issues, online algorithms, data visualization, streaming and on the top analytics of various types of big data. Figure 2 provides the big data modelling methodology.

![Big Data modelling Methodology](image)

**Figure 2.** Big Data modelling Methodology [25]

5. Result analysis
The analytics of large amount of data streams has been provided through enhanced algorithmic approach to provide the optimal solution for data classification and analytics. The prior labelling is providing the source for real-time analysis. Generally, the data stream analysis is achieved in two ways i.e. offline and online. In first instance there is use of general offline clustering algorithm for data analysis. The case of online analysis works for real time data. The data is being received constantly and instant analytics is performed for taking an inline decision. The algorithms are in place to provide the real time analytics for data generated in streams and for online data [31][32]. The process of clustering is providing the way out for realtime analysis primarily when the data is coming from the social networking sources. These data streams are efficiently used to make the requisite analytics in a specific domain. The big data paradigms use more sophisticated ways that are based on graph computing which includes spectral algorithms. The data visualization and fusion for distributed, large with heterogeneous streaming including fusion is an important issue coming out of data realization using big data. The corresponding analysis of the same can be achieved using genetic algorithm [33].

The big data framework proposed in [9] provides a solution against general ICT based application for using growing use of data, systematic storage, and extraction of domain related datasets and analytics.
The framework starts working with the process of acquisition of data by finding the datasets used for training data models. This process is generally carried out through three best-known approaches. The first one is data discovery which involves searching for new datasets either on web or data lakes [17],[18]. The second approach is data augmentation where datasets are updated using more data from realtime sources. The third approach is data generation in which data is generated through crowdsourcing or by using synthetic datasets.

The next stage is to label datasets which can be considered either by use of existing labels of datasets or crowd-based and sometimes there is also use of weak labels which are less than perfect labels. The labelling can be further categorized on basis of machine learning task and data type. The labels are used in machine learning for purpose of classification and regression [19].

The privacy issues have been addressed by using two main approaches viz k-anonymity and differential privacy. The “k-anonymity” which provides the details of property possessed by some anonymized data [26]. The “differential privacy”, optimizes the accuracy of input queries, which mines through statistical databases and provides a scenario where there are less chances of identifying records. But there are still significant issues pertaining to privacy in which social identification is an important one which pops out when data is merged from variable number of identified sources [27][28]. The services for privacy-preserving with data coming from real-time sources still need attention for building benchmark datasets. This type of data privacy needs to be taken in consideration in variable number of domains for overall securities. A dedicated security system needs to be put in place for sensitive applications and a corresponding education program for technical persons [30]. Figure 3 shows machine learning in smart building environment.

The overall performance of the framework has been performed for classification algorithm performed on big data sets of various domains. The classification accuracy predicted by the proposed model is 98.32% and which provides satisfied results that clusters are getting into the proper domain. The formation of clusters, their classification and at last proper distribution provides the further course of optimal results during analytics process. The algorithm has been provided to initially work on five domains and results are generated as per specified domains for their precision and accuracy of results.

The performance specification of various algorithms can be used as per need of different cluster sets and data. The random forests use minimal attributes and settings are used for default parameters with 90% accuracy [34]. The Neural Networks provides 95% accuracy using human activity recognition data sets [35]. The artificial neural networks on some specified data sets with common features are providing around 92% accuracy [36]. The scalability and performance have been drastically improved in the proposed frameworks.

![Diagram](image-url)
The K-means algorithm initially makes the clusters of data and data evaluation has been enhanced after proper allocation of domains using Data Distribution Algorithm [9]. The best data provided to the developed framework during machine learning process for training the model with accuracy. The dynamic issue of data sets i.e. streams of data from different sources has been addressed through prior domain allocation for stored data. The interoperability among different domains is handled during formation of clusters. Second, many data discovery tools rely on dataset owners to annotate their datasets for better discovery, but more automatic techniques for understanding and extracting metadata from the data are needed. While most of the data collection work assumes that the model training comes after the data collection, another important avenue is to augment or improve the data based on how the model performs. The data cleaning approaches preferred are BoostClean and ActiveClean for improved accuracy. The proper labelling and scalability tradeoffs have been reduced. The diversity in creation of data is a challenge that has been set through programming paradigm by creating libraries and templates. The application specific data has been determined separately for model training with small number of labels. Therefore, empirical techniques have been associated for effectiveness. The Table 1. Shows the overall evaluation of system.

Table 1. Table showing various features of Big Data Systems

| Property            | General Big Data Systems | Big Data Systems with Variable domains |
|---------------------|--------------------------|---------------------------------------|
| File System         | HDFS, FTP, Amazon S3     | Technology Independent                 |
| Computational Time  | Disk Based               | In-Memory and Disk Based               |
| Latency             | High                     | Low                                   |
| Fault Tolerance     | High(Replication)        | Comparatively High                     |
| Data Format         | Structured/Unstructured  | Structured/Unstructured                |
| Efficiency          | Input Dependent          | High                                  |
| Accuracy            | High                     | High                                  |
| Security            | Technology Dependent     | High                                  |

The framework and supported algorithms are generalizing overall process of big data mining and integrating the various domains to collectively work on decision making and future trend generalization [9][11]. The methods incorporated are enhancing the overall accuracy of the system. Table .2 shows the accuracy increment and same has been generalized in Figure.4.

Table 2. Execution Time Reduction

| m(k)(Max. # of items in transactions) | Execution Time(seconds) |
|--------------------------------------|-------------------------|
|                                      | Generalized Database Systems | Big Application Specific Domains | Proposed Framework Evaluation |
|                                      | 20          | 41.89                    | 21.46 | 12.4 |
|                                      | 29          | 64.91                    | 34.27 | 21.01 |
|                                      | 38          | 120.83                   | 62.32 | 45.3  |
|                                      | 47          | 205.82                   | 117.76 | 87.12 |
|                                      | 60          | 524.51                   | 313.23 | 251.5 |
|                                      | 73          | 992.54                   | 627.41 | 530.21 |
|                                      | 85          | 2364.12                  | 1323.71 | 1170.32 |
|                                      | 90          | 0                       | 2654.41 | 2430.24 |
|                                      | 100         | 0                       | 5823.24 | 5463.12 |
The big data execution time is challenge for making application respond timely. The Figure.4 shows reduction of execution time by the big data sets in the framework [9][11]. The novels models implemented comparatively shows best results for the proposed work. The machine learning is helping in making predictions and future decisions.

![Graph showing execution time vs input data](image)

**Figure 4.** Chart showing Accuracy increment

The big data execution time is challenge for making application respond timely. The Figure.4 shows reduction of execution time by the big data sets in the framework [9][11]. The novels models implemented comparatively shows best results for the proposed work. The machine learning is helping in making predictions and future decisions. The results of traditional, general big data and proposed big data framework are shown in Table 3.

| Input Data | Percentage Decrease |
|------------|---------------------|
| Traditional Vs Big Data | General Big Data V/s Proposed Big Data Framework |
| 20         | 48.77               | 42.22             |
| 29         | 47.20               | 38.69             |
| 38         | 48.42               | 27.31             |
| 47         | 42.78               | 26.02             |
| 60         | 40.28               | 19.71             |
| 73         | 36.79               | 15.49             |
| 85         | 44.01               | 11.59             |
| 90         | Big Data only       | 8.45              |
| 100        | Big Data only       | 6.18              |

Table 3. Percentage decrease in carrying processing time
The results showed that there is a limit for traditional systems for acceptance of big data. The multidomain big data framework works better than general systems. In order to reduce the gap between technological generation and growth in creation of data on internet there is need to develop standardized frameworks to keep the process of analytics at par with technology. Figure 5 shows the detailed comparison of metrics for overall performance.

Figure 5. Percentage Decrease in processing time

The various parameters that has been evaluated while considering the overall performance is accuracy, precision, recall and corresponding F-measure. The framework proposed is comparatively showing better results for hybrid algorithm. The hybrid method is compared with various machine learning methods as shown in Table 4.

Table 4. F-measure enhancements for various algorithms

| Methodology       | Accuracy | Precision | Recall | F-Measure |
|-------------------|----------|-----------|--------|-----------|
| ANN               | 0.96     | 0.9593    | 0.9494 | 0.954324  |
| Hybrid Algorithm  | **0.962**| **0.994** | **0.985**| **0.98948**|
| Random Forest     | 0.94     | 0.982     | 0.9853 | 0.983647  |
| SVM               | 0.7824   | 0.765     | 0.7734 | 0.769177  |
| KNN               | 0.924    | 0.942     | 0.954  | 0.947962  |

The accuracy has been measured by considering the process of analytics in hybrid algorithm. The value of accuracy also depends on domain used as there is availability of bulk data for some particular domain which gives actual values than others. But all the domains are showing enhanced results than existing frameworks Figure 6.
The models chosen for the analysis are ANN, Random Forest, SVM, KNN against Hybrid Algorithm. Upon testing of said models the datasets validated their result and authenticity on various metrics. The hybrid algorithm based on random forest comparatively showed best results for general analytics as shown in Figure 7.

The harmonic average which is F-Measure for various algorithms has been evaluated. F-measure is shown below in the Figure 8 for various algorithms in different big data frameworks and for the proposed framework the results are better as compared to existing systems.
Figure 8. F-Measure graph for various algorithms

6. Conclusion and Future Work

The research work presented provides evaluation of functioning of frameworks. The wide use of machine learning and providing of solutions to problems arising out of creation of large amount of data is creating bottleneck for currently available technology. However, the enormous amount of data is also creating opportunities for analytics purpose. The technology in place is not sufficiently handling the process of analytics. Moreover, there are standalone applications for various domains. The research presented evaluates a generalized approach with integrated interoperability of bringing in together various domains. The multi-domain big data framework provides accuracy, speed and secure environment for performing analytics. The multi-domain functionality can be further enhanced using genetic algorithm with seed based storage mechanism. The further management of big data and utilization of services from real time applications can be brushed up by bringing together traditional machine learning, general AI techniques and natural language processing.

References

[1]. Ge M, Bangui H and Buhnova, B Big Data for Internet of Things: A Survey Future Generation Computer Systems 2018
[2]. Liu X and Nielsen P S A hybrid ICT-solution for smart meter data analytics. Energy 115 1710–1722 2016
[3]. Sakurai Y, Matsubara Y and Faloutsos C Mining and Forecasting of Big Time-series Data, In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 919–922 2015
[4]. Dean J and Ghemawat S MapReduce simplified data processing on large clusters Comm of the ACM vol 51 no 1 pp 107–113 2008
[5]. Zaharia M Xin, Wendell R , Das S P, Armbrust T, Dave M, Meng A, Rosen X, Venkataraman J, Franklin S M J et al ApacheSpark a unified engine for big data processing” Comm of the ACMvol 59 no 11 pp 56–65 2016
[6]. Assefi M, Behravesh E, Liu G and Tafti A P Big data machine learning using apache spark MLLib 2017 IEEE International Conference on Big Data (Big Data) Boston MA 2017 pp 3492-3498 doi 101109/BigData20178258338

[7]. Rocklin M Dusk- Parallel computation with blocked algorithms and task scheduling in Proc of the 14th Python in Science Conference no130-136 Citeseer 2015

[8]. Chen M, Mao S and Liu Y Big data a survey Mobile Netw Appl19 (2) (2014) 171–209

[9]. Masroor,D and Munishwar, RA Clustering Hybrid Algorithm for Smart Datasets using Machine Learning, (IJACSA) International Journal of Advanced Computer Science and Applications, Volume 11 No 9 September 2020

[10]. Osman A M S A novel big data analytics framework for smart cities, Future Generation Computer Systems Volume 91, February 2019, Pages 620-633, https://doi.org/10.1016/j.future 201806046

[11]. Masroor D and Munishwar R A Novel Framework for Enhancing QoS of Big Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol 11, No 4, 2020

[12]. Zheng K P, Gao J , Ngiam K Y, Ooi B C and Yip W L J Resolving the bias in electronic medical records, 23rd ACM SIGKDD Conference onKnowledge Discovery and Data Mining; Halifax, Nova Scotia, Canada; Aug 13–17, 2017

[13]. Wainberg M, Merico D, Delong A and Frey B J Deep learning inbiomedicine Nat Biotechnol 2018; 36 829–38

[14]. Le Cun Y, Bengio Y and Hinton G Deep learning Nature 2015;521 436–44

[15]. Grinsven V van M Ginneken B Hoyng C Theelen T Sanchez CFast convolutional neural network training using selective datasampling application to hemorrhage detection in color fundusimages IEEE Trans Med Imag 2016; 35 1273–84

[16]. Haendel M A, Chute C G and Robinson P N Classification ontologyand precision medicine N Engl J Med 2018; 379 1452–62

[17]. Terrizzano I G, Schwarz P M, Roth M, Colino J E Data wrangling The challenging journey from the wild to the lake in CIDER 2015

[18]. Halevy A Y, Korn F, Noy N F, Olston C, Polyzotis N, Roy S and Whang S E Managing google’s data lake an overview of the goods system” IEEE Data Eng Bull vol 39 no 3 pp 5–142016

[19]. Yuji R, Geon H and Steven E W A Survey on Data Collection forMachine LearningA Big Data - AI Integration Perspective DOI 101109/TKDE20192946162 IEEETransactions on Knowledge and Data Engineering

[20]. Mohammad A F, Hayme V T and Chowhan S S A Review of Big Data Environment and Its Related Technologies International Conference On Information Communication And Embedded System(ICICES 2016)

[21]. Wang X S, Xie Y and Luo H Cloud Computing Oriented Retrieval Technology based on Big Data Measuring Technology and Mechatronics Automation (ICMTMA) 2015 Seventh International Conference on IEEE PP 275-278 13-14 June 2015

[22]. Kirti P, Sahare S, Murab A and Sarode M V BIG DATA The Leading Innovative and Productive Framework M M Ghonge International Journal of Scientific Research Engineering & Technology (IJSRET) Volume 2 Issue 12 pp 891-896 March 2014 www.ijsretorg ISSN 2278 – 0882

[23]. Ankik T K, Hemlata C and Surendra Y A Review on Big Data and Its Security Innovations in Information Embedded and Communication Systems (ICIIECS) 2015 International Conference on PP 1 - 5 19-20 March 2015 Same paper

[24]. Sun Y, Chen M, Liu B and Mao S A fault-avoidant routing method for data center networks with regular topology (2013) inn Proceedings of ACM/IEEE symposium on architectures for networking and communications systems (ANCS’13) ACM

[25]. Iqbal R Technological Forecasting & Social Change (2018)https://doi.org/101016/jtechfore 201803024

[26]. Sweeney L K-anonymity a model for protecting privacy Int J Uncertain Fuzziness Knowledge-based Syst 10 (5) (2002) 557–570

[27]. Long N H and Jung J J Privacy-aware framework for matching online social identities in multiple social networking services Cybern Syst 46 (1-2) (2015) 69–83
[28]. Dwork C (2008) Differential Privacy A Survey of Results In Agrawal M Du D Duan Z Li A (eds) Theory and Applications of Models of Computation TAMC 2008 Lecture Notes in Computer Science vol 4978 Springer Berlin Heidelberg https://doi.org/10.1007/978-3-540-79228-4_1

[29]. Veiga J, Expósito R R, Pardo X C, Taboada G L and Tourifio J Performance evaluation of big data frameworks for large-scale data analytics 2016 IEEE International Conference on Big Data (Big Data) Washington DC 2016 pp 424-431 doi 101109/BigData20167840633

[30]. Landau S Educating engineers teaching privacy in a world of open doors IEEE Secur Priv 12 (3) (2014) 66–70

[31]. Cramer K and Singer Y Ultraconservative online algorithms for multiclass problems J Mach Learn Res 3 (2003) 951–991

[32]. Cheng J, Ke Y and Ng W A survey on algorithms for mining frequent itemsets over data streams Knowl Inf Syst 16 (1) (2008) 1–27

[33]. Menendez H D, Camacho D Gany a genetic spectral-based clustering algorithm for large data analysis in IEEE Congress on Evolutionary Computation (CEC) 2015 IEEE 2015 pp 640–647

[34]. Chetty G, White M and Akther F Smart phone based data mining for human activity recognition in Proceedings of the International Conference on Information and Communication Technologies ICICT 2014 3–5 2014 at Bolgatty Palace & Island Resort Kochi India Procedia Comput Sci 46 (2015) 1181–1187 http://dxdoiorg/101016/jprocs201501031 URL http://wwwsciedirectcom /science /article/pij S1877050915000320

[35]. Ronao C A and Cho S B Evaluation of deep convolutional neural network architectures for human activity recognition with smartphone sensors2015

[36]. Why the internet of medical things is the future of healthcare 2017 https://www europeapharmaceuticalreview.com /article/47692/imothealthcare/ [Online; accessed 07-July-2020]

[37]. Ousterhout K, Rasti R, Ratnasamy S, Shenker S and Chun B G Making Sense of Performance in Data Analytics Frameworks Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation May 2015 Pages 293–307

[38]. Narooka P and Choudhary S Performance Evaluation of BigData Analysis with Hadoop in Various Processing Systems” International Journal Of Scientific & Engineering Research Volume 7 Issue 10 October-2016 ISSN 2229-5518

[39]. Priya A C R, Sridhar M Spark An Efficient Framework for Large Scale Data Analytics International Journal of Scientific & Engineering Research Volume 7 Issue 2 February-2016 40IJS 2229-5518

[40]. Shahjaman M, Rashid M M Asifuzzaman MI Akter H Islam SMS Mollah MNH Performance evaluation of different machine learning algorithms in presence of outliers using gene expression data Journal of Bio-Science December 2019 2869-80

[41]. Hong N D, Hao W, Guangquan X, Jiafu W and Muhammad I Big Data Analytics for Manufacturing Internet of Things Opportunities Challenges and Enabling Technologies June 2019 Enterprise Information Systems ISSN 1751-7575 DOI 101080/1751757520191633689

[42]. Fan, Cheng, Xiao, Fu, Yan Chengchu, Liu, Chengliang, Li, Zhengdao, Wang and Jiayuan A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning Applied Energy Elsevier vol 235(C) pages 1551-1560

[43]. Galicia A, Talavera R, Llamas, Troncoso A, Koprinska I and Martinez A F Multi-step forecasting for big data time series based on ensemble learning Knowl Based Syst 163 (2019) 830-841

[44]. Galetsi P, Korina K and Kumar S (2020) Big data analytics in health sector Theoretical framework techniques and prospects International Journal of Information Management 50 206-216 101016/jijinfomgt201905003

[45]. Martinez L, Schwarcz F, Nunez V A and Garcia D V Machine learning classification analysis for a hypertensive population as a function of several risk factors (2018) Expert Systems with Applications

[46]. Chowriappa P, Dua S and Todorov Y Introduction to machine learning in healthcare informatics Machine learning in healthcare informatics Berlin Heidelberg (2014) Springer1-23
[47]. Alkatheri S, Abass S A and Siddiqui M A A Comparative Study of Big Data Frameworks International Journal of Computer Science and Information Security 2018

[48]. Qolomany A A, Benhaddou D and Qadir J (2019) Leveraging Machine Learning and Big Data for Smart Buildings A Comprehensive Survey B Qolomany et al Leveraging Machine Learning and Big Data for Smart Buildings A Comprehensive Survey” in IEEE Access vol 7 pp 90316-90356 2019 doi 101109/ACCESS20192926642

[49]. Franks B, Taming the Big Data Tidal Wave Finding Opportunities in Huge Data Streams with Advanced Analytics Wiley com John Wiley Sons Inc 2012

[50]. Rao B P, Saluia P, Sharma N, Mittal A, and Sharma S V Cloud computing for Internet of Things & sensing based applications in Proceedings of the Sensing Technology(ICST) 2012 Sixth International Conference on IEEE 2012 pp374–380