DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5089

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

INTEGRATING BIG DATA IN CLOUD ENVIRONMENT–A REVIEW

Mr.Deepak Ahlawat
PhD Research Scholar MMU Sadopur
Ambala Haryana, India

Dr.Deepali Gupta
HOD CSE MMU Sadopur
Ambala Haryana, India

Abstract: In this paper the concept of the Big Data and Cloud Computing are integrated and reviewed. Big data term refers to huge volume of data in today's internet environment, much of which cannot be integrated easily. Cloud computing and big data go hand in hand. Big data gives the users the ability to utilize massive computing power to process the distributed queries in different datasets and return outcome sets in a timely manner. Cloud computing is the paradigm on which various resources are spread and with the use of Hadoop these can be utilized efficiently. Furthermore, the future work of the integration of big data and cloud computing paradigm are also presented.

Keywords: GA, PRF, CURE.

1. INTRODUCTION

1.1. Big Data

Big data [1] can be characterized by 4Vs: the extreme volume of data, the wide variety of types of data, the velocity at which the data must be must processed and the value of the process of discovering huge hidden values from large datasets with various types and rapid generation. . Big data term refers to huge volume of data in today's internet environment, much of which cannot be integrated easily.

Big data takes huge amount of time and costs/money to get some useful analysis done on it. As knowledge can only be drive from a careful analysis of data (Data Mining), thus several new approaches to storing and analysing data have emerged. Instead, raw data with extended metadata is aggregated in a data lake and machine learning and artificial intelligence (AI) programs use complex algorithms to look for repeatable patterns [2].Collection of large amount of data takes place because of the human involvement in the digital space. The work is being shared stored and managed and lives online. As an example, approximately several terabytes of data daily uploaded and viewed on Facebook.

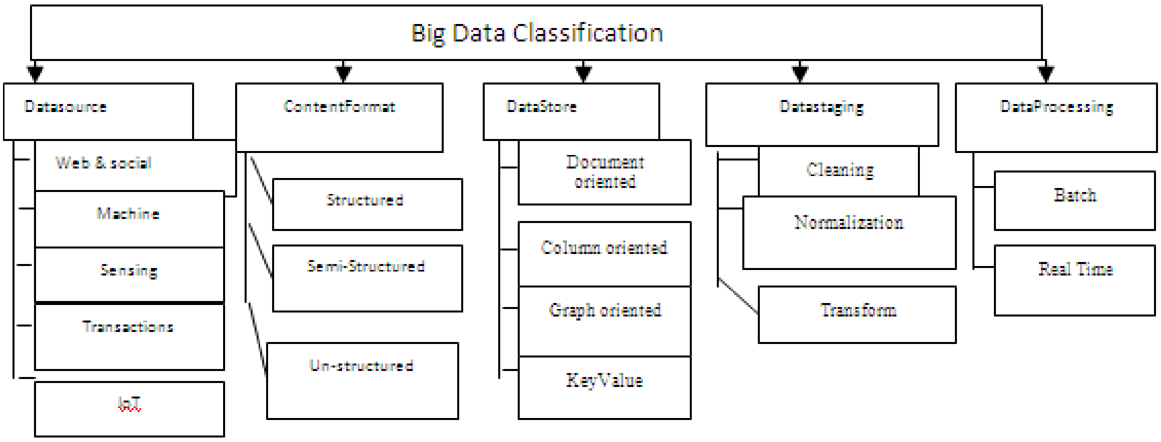

Fig.1. Big Data Classification

This kind of huge data with useful information is known as big data. Clustering is the capable data mining method using widely for mining valuable information in the unlabeled data. From the last few decades, numbers of clustering algorithms are developed on the basis of a variety of theories plus applications.

1.2. Cloud Computing

A cloud is a computing process in which services are dispersed above network by computing processes [3]. Service models consist of three main categories [4]:

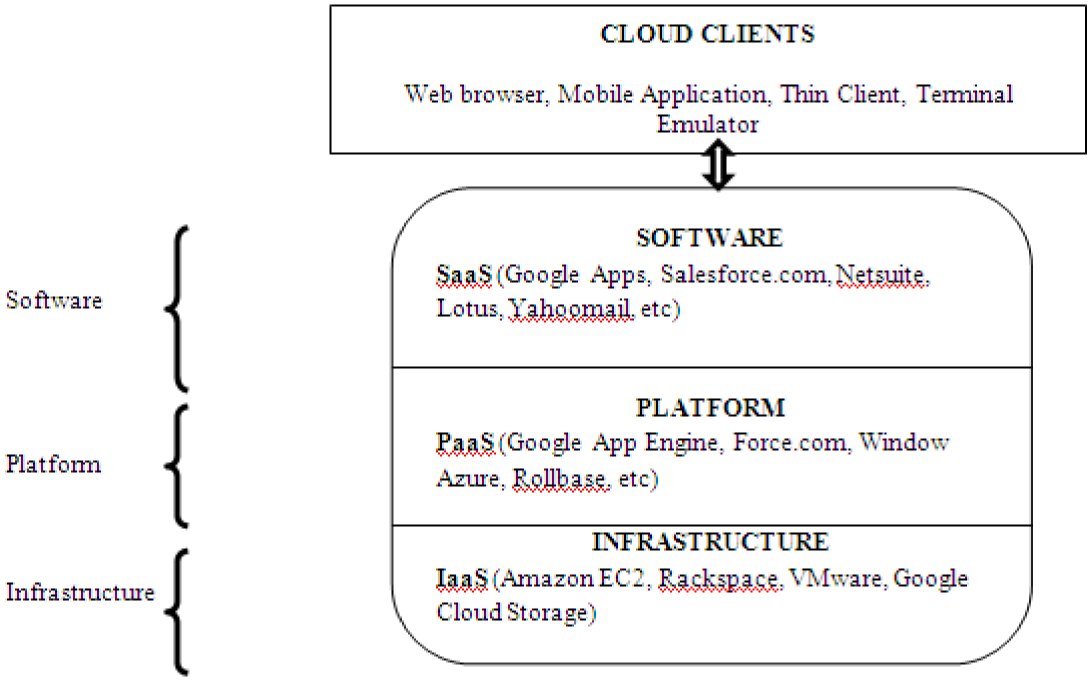
CLOUD CLIENTS
Web browser, Mobile Application, Thin Client, Terminal Emulator
SOFTWARE
SaaS (Google Apps, Salesforce.com, Netsuite, Lotus, Yahoomail, etc)
PLATFORM
PaaS (Google App Engine, Force.com, Window Azure, Rollbase, etc)
INFRASTRUCTURE
IaaS (Amazon EC2, Rackspace, VMware, Google Cloud Storage)
Software
Platform
Infrastructure

Fig.2. Service Models

SaaS (Software as a Service)

- The web access is given to commercial software.
- From a middle location, the software is managed.
- One –to-many is the way for delivering the software.
- The users don’t need to manage software improvements and patches.
- Among number of software’s, Application Programming Interfaces (APIs) allows the integration.

PaaS (Platform as a Service)

- To allow the services to expand, experiment, organize, host and protect the application in the same integrated improved atmosphere and the equivalent services desired to accomplish the application development procedure.
- The web build user interface formation tools assists to make, adapt, test and organize dissimilar UI framework.
- Multi-tenant plan that has numerous simultaneous users use the similar growth application.
- Constructed in scalability of deployed software counting load balancing and failover.
- Addition with the web services and databases of frequent standards.
- Sustain for growth team collaboration – some PaaS solutions comprises of project planning and communication tools.
- Tools to handle billing and subscription management.

IaaS (Infrastructure as a Service)

- The resources are dispersed as a service.
- It permits for effectual scaling.
- It has a patchy cost, usefulness pricing model.
- Usually it has a multiple user environment.

1.3. Relation of Cloud Computing and Big Data

In today’s computing world, most of the software companies don’t provide the complete setupfiles of the software’s, instead the downloading process and installation process goes over the Internet. This type of scenario is possible only through the concept of cloud computing. Huge volume of data is present on these clouds thus; there must be common methods which should be taken into account whenever addressing clouds and big data. The relation between cloud computing and big data [5] is shown in below figure. The huge volume of data or big data is present on clouds which can be accessed via the programming methods that are hidden from a naïve user [6]. With Hadoop, one can easily access and make use of the various resources in the integrated environment.

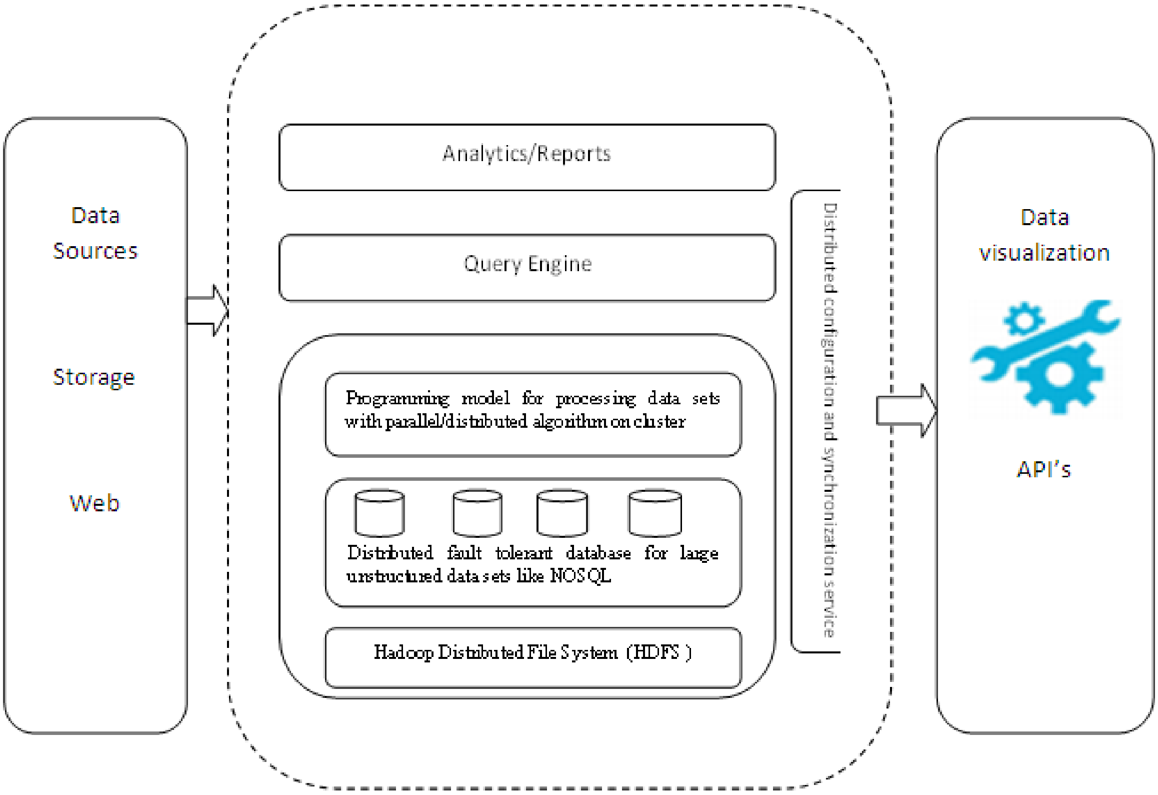
Data
Sources
Storage
Web
Analytics/Reports
Query Engine
Programming model for processing data sets with parallel/distributed algorithm on cluster
Distributed fault tolerant database for large unstructured data sets like NOSQL
Hadoop Distributed File System (HDFS)
Distributed configuration and synchronization service
Data
visualization
API's

Fig.3. Cloud Computing and Big Data

1.4. Clustering in Big Data

Data clustering is known as a problem of a partition of unlabeled objects sets that is $O = \{o1, o2, . . . , on\}$ in k groups of alike objects, in which $1 < k < n$. Before clusters could be required, it become necessary for estimating *k*, this is the problem of cluster tendency [6]. When each object is shown by attributes vector, data clustering is developed on feature vectors **xi** ϵ R**p**, in which **xi** is the *p*-dimensional feature vector for *oi*, $1 \leq i \leq n$. These data could be shown in the form of an $n \times n$ dissimilarity matrix *D*, having *Dij*representing dissimilarity (distance) among *oi*plus *oj*. Basically, the Euclidean distance ||**xi** − **xj**|| is known as the dissimilarity measure, but it could be some norm on R**p**[7].

Following are some of the clustering algorithms:

1.4.1. K-mean clustering

The k-means clustering algorithm is the fundamental algorithm which is dependent on the partitioning method using for many clustering tasks mainly with low dimension datasets. It utilizes k as a parameter, with the division of n objects in k clusters for the objects in the similar cluster to behave similar to every, but different to another objects in other clusters. The algorithm normally finds the cluster centers, (*C1* ...… *Ck*), for minimizing the sum of the squared distances of every data point, *xi*, $1 \leq i \leq n$, *to* its nearest cluster center*Cj*, $1 \leq j \leq k$. Initially, the algorithm arbitrarily selects the k objects, showing a cluster mean/center. Later, the object *xi* in the data set is transferred to the adjacent cluster center i.e. to the parallel center. The algorithm calculates the novel mean for every cluster and re-assigns every object to the adjoining new center. This method iterates till no amendments occur for the assigning the objects. The convergence outcome minimizes the sum-of-squares error which is defined as the squared distances sum from every object to its cluster center [7].

1.4.2. Fuzzy K-mean

Fuzzy K-Means is also known as Fuzzy C-Means Clustering, which is the extension of K-Means technique [8]. The K-Means algorithm only finds the clusters of regular shapes, i.e., Hard Clusters, but Fuzzy K-mean is also suitable to find the Soft Clusters [9].

The fuzzy k-means algorithm is described as follows:

1. To assume a fixed number of clusters *k*. To Randomly initialize the k-means μ_k connected with the clusters with the computation of the probability that every data point $x_i$is a member of a known cluster *k*,

$$P(pointx_i\ has\ labelk|x_i, k)$$

2. To recalculate the centroid of the cluster as the weighted centroid mentioned the probabilities of membership of all data points x_i:

$$\mu_k(n+1) = \frac{\sum_{x_{i\epsilon k}} * P\ (\mu_k|x_i)^b}{\sum_{x_i \epsilon k} P\ ((\mu_k|x_i)^b}$$

3. To iterate till convergence of a user-specified number of iterations being reached.

1.4.3. Clustering using Genetic Algorithm

GA (Genetic algorithm) was proposed early in 1989 that attracts many attentions as it perform a globalized investigation for solutions whereas another clustering approaches execute a localized search and therefore, simply get stuck at local optimality's. In a localized search, the novel obtain solution take over the ones in the preceding iteration. Such example includes k-means, ANNs, fuzzy clustering algorithms with tabu search,annealing schemes. However, in Genetic Algorithm, the crossover and mutation operators could produce novel solutions that are very dissimilar from the preceding iteration which is where the global optimality basically comes [10]. Also, Genetic algorithm works paralleling, making it possible for implementing parallel hardware for speed up the execution.

In fact, Genetic Algorithm is known as evolutionary approach, which applies evolutionary operators and solutions population for achieving a partition of global optimal. GA includes selection of functions, mutation operation,and a fitness function. The candidate solutions to the clustering problem are being encoded as chromosomes, and later a fitness function inversely proportional to the squared error value is applied for determining the chromosomes existing likelihood in the subsequent generation [11].

2. RELATED WORKS

Chen et al., (2017) optimized the previous research results which are implemented on Spark MLib. The author makes use of hybrid technique on Parallel Random Forest (PRF) algorithm. The author integrate the data-parallel optimization and task-parallel optimization [12].

Xu et al., (2017) had designed a speculative execution schemes for parallel processing clusters. Researchers devised two schemes: one for lightly loaded systems and other for heavily loaded systems. For light loaded systems, they proposed Smart Cloning Algorithm (SCA) and for heavily loaded systems, Enhanced Speculative Execution (ESE) Algorithm is proposed. The experimental result of the SCA and ESE algorithm are compared with Microsoft Mantri.The SCA has reduced the job flowtime by 6% in comparisonto Microsoft Mantri. In terms of the job flowtime, the ESE algorithm outperforms the Microsoft Mantri baseline scheme by 71%[13].

Thingom et al., (2017) discusses the concept of the integration of big data and cloud computing. Researchers pointed out the flexibility and minimum cost (pay & use model) required in the cloud scenario [14].

El-Seoud et al., (2017) showed the trends and challenges faced in the field of big data and cloud computing. Study reveals the risks plus benefits that may arise due to the integration of big data and cloud computing. The study also unfolded the concepts behind big data and cloud computing [15].

Bharill et al., (2016) has focused his paper on clustering large datasets in Apache Spark environment. Authors designed and implement partitioned dependent clustering and choose the specified environment because of its low computational needs. In this research, Scalable Random Sampling with Iterative Optimization Fuzzy C-Means algorithm (SRSIO-FCM) is implemented on an Apache Spark Cluster. The experimental studies on different big datasets are conducted. The performance of SRSIO-FCM is better in comparison to the Literal Fuzzy C-Means (LFCM). The results are stated in terms of space and time complexity. According to the results, SRSIO-FCM runs in less time without compromising the clustering quality [16].

Wei Shao et al., (2016) have presented a model for clustering data by means of spatiotemporal-intervals, which is consider as a spatiotemporal data type connected with a start- and an end-point. The model proposed by the researcher could be used to evaluate the spatiotemporal interval data clusters. The work has aimed to deal with the evaluation of clustering results in variety of Euclidean spaces. This is dissimilar from the existing clustering that calculates the outcome in space of single Euclidean. The existing clustering algorithms are analyzed and compared with the use of energy function [17].

Sun et al., (2015) has done clustering with the use of time impact factor matrix. The matrix monitors how user interest drifts and then predicts the rating of the item.In addition to the time impact factor matrix, the author has added one more time impact factor and use the linear regression for predicting the user interest drift. The comparisons of the experiments have been conducted on three big data sets, namely, MovieLens1M, MovieLens100K, and MovieLens10M. The results have shown that the proposed approach has efficiently improved the prediction accuracy [18].

Sookhak et al., (2015) improves the storage capability of the cloud system by reducing the communicational and computational overhead costs. Authors proposed an RDA (Remote Data Auditing) technique which is dependent on algebraic signature properties. The authorsproposed DCT (Divide and Conquer Table) which is a data structure that could perform the operations such as, insert, delete append or modify. The comparison among the proposed method and other previous RDA techniques has shown that the proposed method is more secure and also reduces the computational and communicational overhead [19].

Kumar et al., (2015) proposed the ClusiVAT algorithm. The proposed algorithm is compared with the K-means, single pass K-means, online K-means andCURE (Clustering using representatives).The comparison results show that ClusiVATis the fastest and accurate among all five algorithms. For example, it has recovered 97% of the ground truth labels in the real world KDD-99 cup data (4 292 637 samples in 41 dimensions) in 76 s [7].

Hashem et al., (2014) studied how the vast amount of data (Big data)and cloud computing is a challenge in today's computer world. Author discusses all the issues regarding Big Data in Cloud Computing environment. Furthermore, the Latest research challenges are also addressed [1].

Yin et al., (2014) have focussed on the detection of faults with the isolation for the systems of vehicle suspension. The system being proposed is classified into mainly three steps, primarily to confirm the number of clusters dependent on PCA (Principal component analysis and secondly to detect the faults by using fuzzy positivistic C-means clustering with the fault lines and next to isolate the root causes for faults by using the technique of Fisher discriminant analysis. Dissimilar from another scheme, the proposed method only requires measurements of accelerometers which are fixed on four corners of a vehicle suspension. Moreover, dissimilar spring attenuation coefficients are being regarded as a special failure in place of few others [8].

Konak et al., (2006) studied the emerging technology GA (Genetic Algorithm) for the existing problems.Author addresses the multi-objective formulations which are considered as realistic techniques for problems of more complex engineering optimization. For real-life problems, the objectives under consideration conflicts with each other and the optimization of particular solution for single objective that could result in unacceptable results for other objectives [11].

3. FUTURE WORKS

Beyond the basic execution needs, small additional services like Machine learning, Analytics, and Orchestration are being accessible by the cloud. There are numerous reasons for this move as summarized below [20]:

i. Clouds are the main providers for data services.
ii. Machine Learning and other AI approaches will surely improve the scenario and Orchestration (Automation) would make the service provider capable to have the Service level agreement on time.
iii. Analytics would accelerate the business and Orchestration can be helpful when the acceleration takes place.
iv. The future of Clouds would be the mixture of Analytics and Orchestration.
v. Big Data and Cloud Computing will surely automate the maximum workload in the distributed computing environment.

REFERENCES

1. Hashem et al., "The rise of “big data” on cloud computing: Review and open research issues”, Information Systems 47 (2014): 98-115.
2. Wu et al., “Data mining with big data”, IEEE transactions on knowledge and data engineering 26.1 (2013): 97-107.
3. Subashini et al., “A survey on security issues in service delivery models of cloud computing”, Journal of network and computer applications 34.1 (2011): 1-11.
4. Pallis et al., “Cloud computing: the new frontier of internet computing”, IEEE internet computing 14.5 (2010): 70-73.
5. Talia Domenico, “Toward cloud-based big-data analytics”, IEEE Computer Science (2013): 98-101.
6. Fernandez et al., “Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4.5 (2014): 380-409.
7. Kumar et al., “A hybrid approach to clustering in big data”, IEEE transactions on cybernetics 46.10 (2015): 2372-2385.
8. Yin et al., “Performance monitoring for vehicle suspension system via fuzzy positivistic C-means clustering based on accelerometer measurements”, IEEE/ASME Transactions on Mechatronics 20.5 (2014): 2613-2620.
9. Zahid et al., “Fuzzy clustering based on K-nearest-neighbours rule”, Fuzzy Sets and Systems 120.2 (2001): 239-247.
10. Maulik et al., “Genetic algorithm-based clustering technique”, Pattern recognition 33.9 (2000): 1455-1465.
11. Konak et al., “Multi-objective optimization using genetic algorithms: A tutorial”, Reliability Engineering & System Safety 91.9 (2006): 992-1007.
12. Chen et al., “A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment”, IEEE Transactions on Parallel and Distributed Systems 28.4 (2017): 909-933.
13. Xu et al., “Optimization for Speculative Execution in Big Data Processing Clusters”, IEEE Transactions on Parallel and Distributed Systems 28.2 (2017): 530-545.
14. Thingom et al., “An Integration of Big Data and Cloud Computing”, Proceedings of the International Conference on Data Engineering and Communication Technology (2017): 729-737.
15. El-Seoud et al., “Big Data and Cloud Computing: Trends and Challenges”, International Journal of Interactive Mobile Technologies 11.2 (2017): 34-52.
16. Bharill et al., “Fuzzy Based Scalable Clustering Algorithms for Handling Big Data Using Apache Spark”, IEEE Transactions on Big Data 2.4 (2016): 339-352.
17. Wei Shao et al., “Clustering Big Spatiotemporal – Interval Data”, IEEE Transactions on Big Data 2.3 (2016): 190 – 203.
18. Sun et al., “Dynamic Model Adaptive to User Interest Drift Based on Cluster and Nearest Neighbors”, IEEE Access 14.8 (2015): 1682-1691.
19. Sookhak et al., “Dynamic remote data auditing for securing big data storage in cloud computing”, Information Sciences 380 (2015): 101-116.
20. Furht et al., “Handbook of cloud computing”, Vol. 3. New York: Springer, (2010).
21. Azar et al., “Dimensionality Reduction of Medical Big Data using Neural-Fuzzy Classifier”, Soft Computing: Springer 19.4 (2015): 1115-1127.
22. Cao et al., “Cluster as a Service: A Resource Sharing Approach for Private Cloud”, Tsinghua Science and Technology 21.6 (2016): 610-619.
23. Han et al., “Data Mining: Concepts and Techniques”, Elsevier (2006).
24. Kurasova et al., “Strategies for Big Data Clustering”, IEEE 26th International Conference on Tools with Artificial Intelligence (2014): 740-747.
25. Reshmy et al., “Data Mining of Unstructured Big Data in Cloud Computing”, International Journal of Business Intelligence and Data Mining 12.3 (2017).
26. Zhao et al., “Independent Tasks Scheduling Based on Genetic Algorithm in Cloud Computing”, WiCom ’09- 5th International Conference (2009).