International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 09 | Sep 2019 www.irjet.net p-ISSN: 2395-0072

Comparatively Analysis on K-Means++ and Mini Batch K-Means Clustering Algorithm in Cloud Computing with Map Reduce

Virendra Tiwari[1], Dr.Akhilesh A. Waoo[2]

[1]*Assistant Professor, Dept. of Computer Science & Application, AKS University, Satna, Madhya Pradesh, India*
[2]*Associate Professor, Dept. of Computer Science & Application, AKS University, Satna, Madhya Pradesh, India*

Abstract - *Cloud computing provides the functionality to host the servers in a dispersed network so wide range of customers can access the applications as utilities over the internet. One of the very important aspects for the development of cloud computing is the rapidly growing the volume of Big data which requires to be managed and examined efficiently. Experts and Researchers design and utilize multiple algorithms with intelligent approach to gain useful knowledge from a huge volume of data set. Hadoop MapReduce is a new and emerging software platform suitable for writing applications easily which process huge amounts of data in-parallel on large clusters of commodity hardware in a reliable and fault-tolerant manner. There only specific unsupervised learning algorithms for big data problems which are working and being executed successfully in MapReduce technique and are deployed on high volume of data set. MapReduce frameworks performance suffers degradation due to the sequential processing approaches. To overcome this limitation and reducing costs, we can introduce cloud computing with parallel processing in MapReduce a powerful and effective approach for the future technology enhancements. This survey gives a brief overview of cloud computing and comparatively analysis on most popular clustering algorithms named K-means++ and Mini Batch K-means.*

Key Words: Cloud computing, Big data, Hadoop, MapReduce, K- Means++, and Mini Batch K-means.

1. INTRODUCTION

Cloud computing in simple terms on demand availability of hardware resources for computing or storage purpose over the Internet instead of our computer's hard drive. With cloud computing, clients can access files and many applications from any remote device through internet. It reduces the work and cost for client to setup and manage the hardware and required infrastructure [11][12]. Cloud provider take the ownership of providing the set up and maintaining the hardware infrastructure. Most famous cloud service providers are Amazon, Alibaba, Google and Microsoft [13].

Most famous Cloud Service Providers [11] [14]

a) **Google**: Google Cloud Platform solves issues with accessible AI & data analytics and uses resources such as hard disks, computers, virtual machines, etc. located at Google data centers. Google is a dedicated cloud computing service provider, with all the storage available online so it can work with the cloud apps: Google Sheets, Google Docs, and Google Slides. Most of Google's services could be considered cloud computing: Gmail, Google Maps, Google Calendar, Picasa, Google Analytics and so on.
b) **Alibaba**: Alibaba is one the largest cloud computing platform which created a global footprint with over 1500 CDN Nodes worldwide of 19 regions and 56 availability zones across more than approximate 200 countries. Its prominent features are helping to protect and backup your data and help you to achieve faster results.
c) **Apple iCloud**: Apple's cloud service is mainly used for online backup, storage, and synchronization of your mail, calendar, contacts, and more. All the data we need remains available on Mac OS, iOS, or Windows device.
d) **Amazon Cloud Drive**: Web hosting platform which storage reliable and cost-effective solutions. If you have Amazon Prime, you get unlimited image storage. It's essentially storage for anything digital and It is the most popular as it was the first to enter the cloud computing space.

World has created 90% of the data in the last two years which comes from everywhere and surely this is only going to grow with arrival of internet of things. By 2020, it's estimated that approx 1.7MB of data will be created every second for every person on planet"[13]. Processing such a huge amount of data will be extreme challenging for the data analyzers and researchers. Classical sequential methods of programming are not very efficient and hence the requirement of parallel processing gained importance. Cloud provides the most suitable environment and Hadoop MapReduce as the processing components a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster [16][17].

Virtually we are floating in oceans of data, with the advent of Big data; MapReduce emerged as a most effective solution. Number of researchers work is undergoing on the extension of MapReduce carried out with new features, functionalities and mechanism to perfect it for a new set of problems [18].Various sequential programming algorithms are nowadays being use to convert to Hadoop MapReduce programming paradigm. There may be few algorithms that cannot be parallelized, so their uses due to working efficiency are less in the real world as parallelization is of essence to process Big Data. Hence some of the conventional useful algorithms for clustering may provide a great

platform and enhancements to the new emerging technology and also very beneficial to the business enterprise organizations [13].

2. BACKGROUND

a) Cloud computing

Cloud computing is the delivery of different services like managing hardware and software through the Internet. It has evolved since the inception of internet due to its efficiency in storing data, computation and less maintenance cost. Cloud-based storage is making it possible to save files and documents to a remote database and retrieve them on demand. Services can be in public and private. Public services are providing online for a fee while private services are being hosted on a network to particular clients [19].

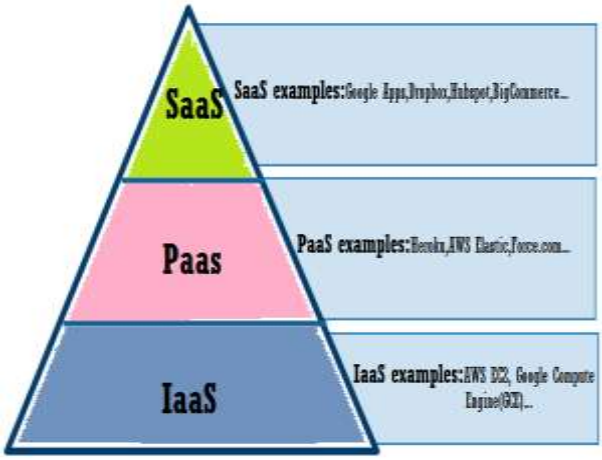

Figure. 1: Cloud computing services type

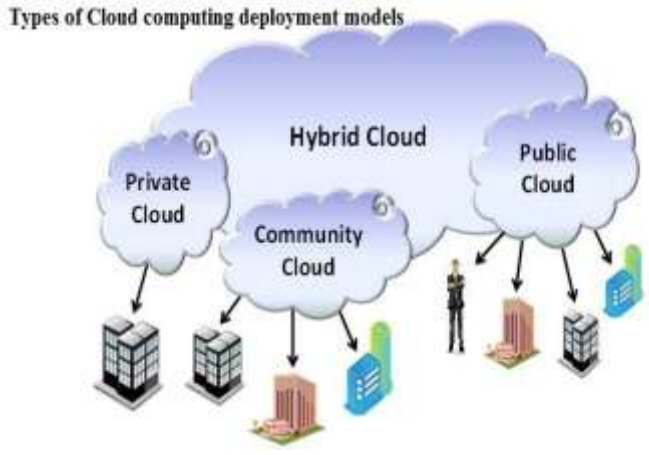

Figure. 2: Cloud computing deployment models

As shown in Fig. 1, cloud providers services can be categorized in three delivery types that are Software as a Service (SaaS), Infrastructure as a Service (IaaS) and Platform as a Service (PaaS). Cloud deployment services comes in the form of private, hybrid, public and community based on their location as mentioned in Figure 2.Public clouds based on a shared cost model have many customers compared to other forms of cloud. Private clouds are perfect for organizations that have high management demands, high-security requirements, and availability requirements usually but are not cost efficient. Community Cloud mutually shared model between organizations [21].

b) Hadoop

Hadoop is an open-source software framework considered to be one of the best tools to handle and analyze very huge volume of data. Currently being used by Google, Facebook, LinkedIn, Yahoo, Twitter etc. It has two major components Hadoop Distributed File System (HDFS) and another is MapReduce. Hadoop is being highly in used distributed Computing platform specially in past 10 years as big data is getting evolved with social media generating PETA bytes of data everyday which can be used for data mining, predictive analytics, and machine learning applications. Hadoop can run with single node or multi-node cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment. When HDFS takes data as input, it breaks the information down into various separate blocks and distributes all the sepreted blocks to different nodes in a cluster; it provides highly efficient parallel processing. HDFS can be thought of Data node (stores actual data in HDFS) + Name node(holds the meta data for the HDFS) + Secondary Name node(merging editlogs with fsimage from the namenode) and daemon process to manage MapReduce programming paradigm in HDFS are Job Tracker[20][13].

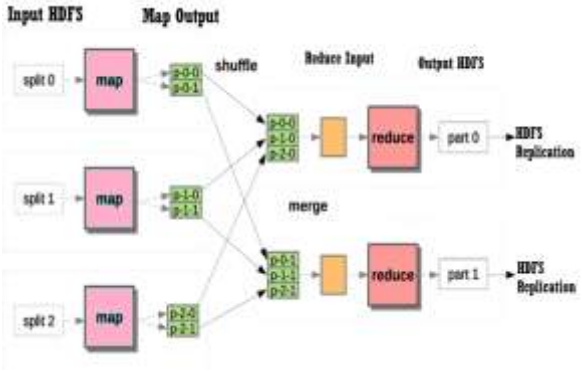

Figure. 3: Data flow in HDFS

c) Map reduce paradigm

Map Reduce Programming Paradigm technique and a program model for distributed computing to process huge amount of data. MapReduce program executes in three stages: Map phase process and maps the input and creates several small chunks of data into key-value pairs [3][22].Complete process gets executed by four phases namely splitting, mapping, shuffling, and reducing. The reduce phase combines the data based on common keys and performs reduce operation defined by the user. The parallelization implementation occurs with many mappers created for reading the data and it is not sequential. Because of this there is high throughput [5][21].The core advantage

of the MapReduce framework is its fault tolerance due to periodic reports from each node in the cluster are expected when work is processed.

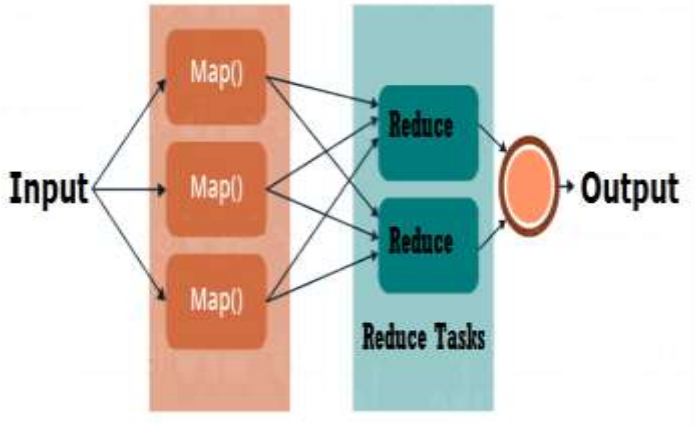

Figure. 4: Map reduce in HDFS

3. K-Means++ CLUSTERING ALGORITHM

K-Means++ algorithm is the improved and robust version of most popular and widely used K-means algorithm in clustering. K-Means algorithm chooses and initializes the centroid randomly, forms the clusters for a given dataset [23].

The performance and accuracy of k-means depends completely on the goodness of the initialization where this algorithm observed and found sometimes giving less accurate clusters. To overcome this limitation, we can use improved clustering method K-Means++ which overcomes by initialization part to improve K-Means algorithm [13][24]. K- Means++ technique, takes an input k, which refers to the number of clusters that should be generated and n refers to set of objects [4]. This improved algorithms mainly aims to initialize the centroids for traditional k-means clustering algorithm [25].K-Means++ algorithm reduces the computational time and it can handle huge amount of data sets.

K-Means++ clustering is one of the partitioning based clustering algorithm works as follows [25].

[1] Choose initial centroid X uniformly at random from initial data set.
[2] For each data point X we need to compute D(X) which is the minimum squared distance between X and the nearest centroid that has been already defined.
[3] Choose another data point as second centroid using a weighted probability distribution which is proportional to D(X).
[4] Repeat 2 and 3 until required K centroids have been chosen.
[5] Assign each record or instance point to a cluster centre which has least distance.
[6] Replace cluster centroids by calculating mean value of all points in the cluster.

In map-reduce pattern mentioned algorithm can be implemented as follows.

- **Map function**: The HDFS stores input data as sequence file of <key, value> pairs and produces a set of <key, value> pairs as the output of the job[6]. The map function splits the data across all mappers, creates a map task for each split and assigns each map task to a Task Tracker[7].First Step is Map, Second Step is Shuffle and Third Step is Reduce[26][27].
- **Reduce function**: After mapping, reducers are used for computing Step-2 of the mentioned algorithm. The Reducer's job is to process the data that comes from the mapper [8]. After processing, the intermediate data can be put in hdfs or stored locally [9]. Afterwards new centers will be generated which can be used for further iterative statements [27].

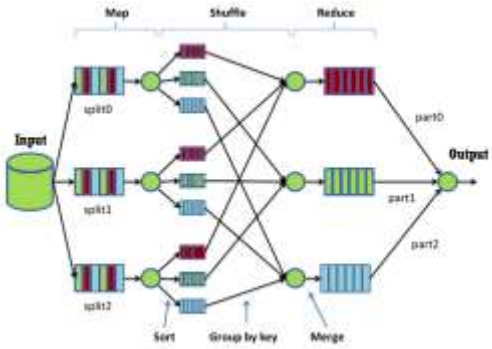

Figure. 5: K-Means++ Map Reduce

4. Mini Batch K-means

The mini batch k-means clustering algorithm [29] is the alternative and modified version of the k-means algorithm. The advantage of this algorithm uses mini-batches to reduce the computation time and cost in large datasets not using all the dataset each iteration [33]. Moreover, it attempts to provide optimal result of the clustering. To accomplish this, the mini batch k-means takes mini-batches of a fixed size as an input randomly, which are subsets of the all dataset [30].

The Mini Batch K-Means framework with Map-Reduce [31][32]:

1. From the throughput of the input sequence, initial cutoff gets calculated using a sample of points.

2. Here, MapReduce job splits the input files into chunks and further assign each split to a mapper to process that consists of an instance of the k-means algorithm this is called InputSplit.

3. It will be giving a set of centroids c that is much larger than the final expected clusters k but, it should be small enough to fit into the memory.

4. Now, Each split which is the results of clustering gets pass to single reducer uses the streaming Ball k-means function to combine the clusters as required.

Before invoking the Map routine, OutputFormat organizes the key-value pairs from Reducer for writing it on HDFS. Based on optimal distance-cutoff parameters are required to measure the distance between the point and the centroids, Either the point is merged to any of the existing clusters or become a new centroids to form new cluster. To determine Optimal Clusters, the points which are forming the clusters with any of centroids become the value in key-value pair.

5. CONCLUSIONS AND FUTURE WORK

Today, Big Data and Cloud Computing are the two mainstream technologies to managing and processing big data Big data makes you dealing with the massive scale of data whereas Cloud computing provides you infrastructure. Today Big Data generating the demand of efficient computing, well managed storage resources as well as best way to process the data [27]. Conventional methods are not found by the researchers and failing to cope up with such challenges because of which Cloud Computing and Hadoop MapReduce found efficient and gained popularity [10]. MapReduce is intelligent enough to tackle huge data sets by distributing processing across many nodes, and then reducing or combining the results of those nodes. In summary, Mini Batch K-means and K-Means++ algorithms discussed and analyzed in this research are robust and efficient for large set of data. K-means++ always tries to select centroids that are far away from the existing centroids, which gives significant improvement over others with the issue of outlier whereas Mini Batch K-means algorithm is not only found suitable and efficient for the processing of huge data, but also make sure the accurate by reduce the amount of computation required [28][31]. It also found giving relatively stable precision, for the outlier detection of data. For scaling well and efficiently process large datasets this algorithm requires to be improved in a way so it could reduce outliers efficiently and to give better and more accurate results with reduced run time.

REFERENCES

[1]Rajashree Shettar, Bhimasen. V. Purohit, “A Review on Clustering Algorithms Applicable for Map Reduce”, International Conference on Computational Systems for Health & Sustainability, pp. 176-178, 2015.

[2]Adem Tepe, GüRay Yilmaz, “A Survey on Cloud Computing Technology and Its Application to Satellite Ground Systems”, International Conference on Recent Advances in Space Technologies (RAST), 2013.

[3]K. Singh and R. Kaur, "Hadoop: Addressing challenges of Big Data," 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, 2014, pp. 686-689.

[4]Weihzong Zhao, Huifang Ma, Qing He, “Parallel K-Means clustering Based on MapReduce”, springer-verlag Berlin, Heidelberg, 2009.

[5]Borthakur. D “The Hadoop Distributed File System: Architecture and design”, 2007.

[6]Sangeeta Ahuja, M.Ester, H. P. Kriegel, J. Sander, X. Xu, “A Density based algorithm for discovering clusters in large spatial database with noise”, Second international conference on knowledge discovery and Data Mining,1996.

[7]B. Dai and I. Lin, "Efficient Map/Reduce-Based DBSCAN Algorithm with Optimized Data Partition," 2012 IEEE Fifth International Conference on Cloud Computing, Honolulu, HI, 2012, pp. 59-66.

[8]V. Gaede, O. G’unther, “Multidimensional access methods”, ACM comput. Surv., Vol. 30, No. 2, pp. 170- 231, 1998.

[9]Varad Meru “Data clustering: Using MapReduce”, software Developers Journal, 2013.

[10]Das A.S, Datar M, Garg, and Rajaram S, “Google news personalization scalable online collaborative filtering”, pp. 271-280, 2007.

[11]Aaqib Rashid and Amit Chaturvedi, "Cloud Computing Characteristics and Services", 2019.

[12]Mahantesh N. Birje and Praveen S. Challagidad "Cloud computing review: concepts, technology, challenges and security “, 2017.

[13]Sweekruth S Badiger and Sushmitha N,"A Review on K-means++ Clustering Algorithm in Cloud Computing with Map Reduce”, 2019.

[14]Dr. Richa Purohit,"Comparative Analysis of Few Cloud Service Providers Considering Their Distinctive Properties”, 2017.

[15]Vishal R. Pancholi and Dr. Bhadresh P. Patel,"A Study on Services Provided by Various Service Providers of Cloud Computing", 2017.

[16]https://www.ibm.com/analytics/hadoop/mapreduce.

[17]Dr. Anitha Patil,"Securing mapreduce programming. Paradigm in hadoop, cloud and big data eco-system", 2018.

[18]Jeffrey Dean and Sanjay Ghemawat,"MapReduce: Simplified Data Processing on Large Clusters", 2004.

[19]Palvinder,"Survey Paper on Cloud Computing”, 2014.

[20]Nasrullah and Ms. Akanksha Bana, "Review paper on hadoop configuration and implementation in virtual cloud environment", 2018.

[21]Santhosh voruganti,"Map Reduce a Programming Model for Cloud Computing Based On Hadoop Ecosystem", 2014.

[22]Bisma Bashir, "An Approach of MapReduce Programming Model for Cloud Computing", 2017.

[23]Md. Zakir Hossain, Md. Nasim Akhtar, R.B. Ahmad and Mostafijur Rahman,"A dynamic K-means clustering for data mining ", 2019.

[24]Dhruv Sharma, Krishnaiya Thulasiraman and John N. Jiang, "A network science-based k-means++ clustering method for power systems network equivalence", 2019.

[25]G. Yamini and Dr. B. Renuka Devi, "A New Hybrid Clustering Technique Based on Mini-batch K-means And K-means++ For Analysing Big Data", 2018.

[26]N. Deshai, S. Venkataramana and Dr. G. P. Saradhi Varma3 "Research Paper on Big Data Hadoop Map Reduce Job Scheduling", 2018.

[27]Ch. Shobha Rani and Dr. B. Rama"MapReduce with Hadoop for Simplified Analysis of Big Data", 2017.

[28] Bo Xiao, Zhen Wang, Qi Liu, and Xiaodong Liu "SMK-means: An Improved Mini Batch K-means Algorithm Based on Mapreduce with Big Data", 2018.

[29]Sculley D , "Web-scale k-means clustering", 2010.

[30]Ali Feizollah, Nor Badrul Anuar, Rosli Salleh and Fairuz Amalina, "Comparative Study of K-means and Mini Batch K-means Clustering Algorithms in Android Malware Detection Using Network Traffic Analysis",2014.

[31]Meghana M Chavan, Asawari Patil,Lata Dalvi And Ajinkya Patil, "Mini Batch K-Means Clustering On Large Dataset",2015.

[32] Mr. Krishna Yadav and Mr. Jwalant Baria,"Mini-Batch K-Means Clustering Using Map-Reduce in Hadoop", 2014.

[33]Nadeem Akthar, Mohd Vasim Ahamad and Shahbaaz Ahmad, "MapReduce Model of Improved K-Means Clustering Algorithm Using Hadoop MapReduce", 2016.