International Journal of Research in Engineering, Science and Management
Volume-2, Issue-5, May-2019
www.ijresm.com | ISSN (Online): 2581-5792

A Review on Cloud Computing and K-means++ Clustering Algorithm with Map Reduce

Vinay V. Hegde[1], Nandan Singh Gadwal[2]
[1]*Associate Professor, Dept. of Computer Science and Engg., R. V. College of Engineering, Bangalore, India*
[2]*PG Student, Dept. of Computer Science and Engg., R. V. College of Engineering, Bangalore, India*

***Abstract*: Cloud computing is a concept of hosting servers in distributed centres with an intention to provide access to many customers. Cloud computing has in essence evolved due to the flexibility that it provides to the user for upscaling or downscaling the hardware at any time. Another major reason for the growth of cloud computing is the increasing size of Big Data generated and need for it to be analysed. Many algorithms have been used by researchers to extract useful information from a large data set. Hadoop has come with a new software programming model called as MapReduce for parallel processing of large dataset. Only few unsupervised learning algorithms are successfully implemented in MapReduce fashion and are applied to massive dataset. The flexibility and power of cloud computing combined with parallel processing using MapReduce programming for a unsupervised learning algorithms is thought to be the optimal way going forward. This survey details about cloud computing and one of the famous clustering algorithm.**

***Keywords*: Cloud computing, Hadoop, MapReduce, K-Means++.**

1. Introduction

Cloud computing is on demand availability of hardware resources for computing or storage purpose. The term 'Cloud' basically means available to all through internet. It reduces the work for user to either setup or manage the hardware. Cloud provider take the ownership of providing and maintaining the hardware infrastructure. Most famous cloud service providers are Amazon, Google and Microsoft.

World is producing data of 2.5 quintillion bytes per day at current pace due to internet, and it's only going to increase with arrival of internet of things. Data is new 'oil' of our generation and mining it to discover knowledge is critical to many businesses. Processing such a large amount of data will pose unique kind of challenges to the data analyzers. Normal sequential methods of programming are less efficient and hence the need of parallel processing gained importance. Cloud provides the ideal infrastructure and Hadoop MapReduce programming paradigm can make most use of it.

Many sequential programming algorithms are now converted to Hadoop MapReduce programming paradigm. It should be noted here that there may be algorithms that cannot be parallelized, and hence are of not much use in the real world today as parallelization is of essence to process Big Data. Hence some of the conventional algorithms for clustering or classification are now implemented using MapReduce Programming paradigm.

2. Background

A. *Cloud computing*

Cloud computing has evolved since the inception of internet. Historically processing and storage were expensive, but due to scientific and technological advancement in hardware manufacturing industries, hardware has become much cheaper and smaller in size compared to earlier. Taking advantage of this fact, industries came up with providing the hardware infrastructure as service to users and this is how Cloud computing paradigm was born.

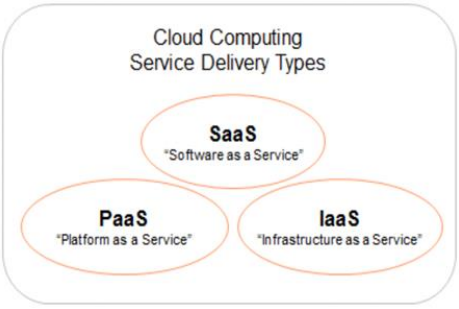

Fig. 1. Cloud computing service types [1]

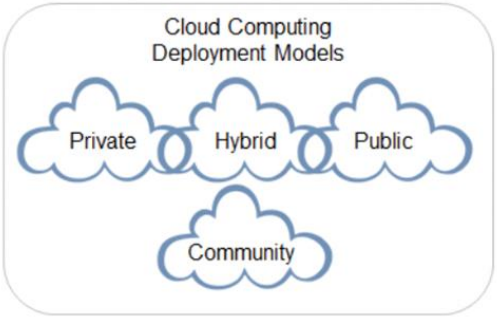

Fig. 2. Cloud computing deployment models [1]

As shown in Fig. 1, cloud providers can provide services in three delivery types, that is Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Cloud deployment services comes in the form of

private, hybrid, public and community as mentioned in Figure 2. Public clouds have many customers compared to other forms of cloud. Private clouds usually are not cost efficient.

B. Hadoop

Hadoop is considered to be one of the best tools to handle big data. It has two major components HDFS and another is MapReduce. HDFS stores the files in blocks of 64MB. It can handle the files of varying size from 10 MB to GB, TB. Hadoop can run with single node or multi-node cluster. Every Hadoop cluster can have five running processes namely. HDFS can be thought of Data node + Name node + Secondary Name node and daemon process to manage MapReduce programming paradigm in HDFS are Job Tracker + Task Tracker.

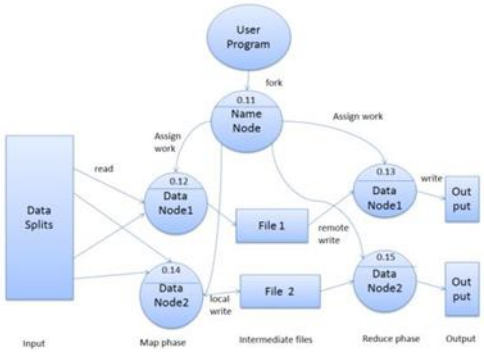

Fig. 3. Data flow in HDFS [2]

C. Map reduce paradigm

Map Reduce Programming Paradigm of Hadoop is a model to process huge amount of data. Map phase maps the input data into <key,value> pairs[3]. The reduce phase combines the data based on common keys and performs reduce operation defined by the user. The parallelization occurs with many mappers created for reading the data and it is not sequential. Because of this there is high throughput [4].

3. K-Means++ clustering algorithm

K-Means++ algorithm is the improved version of most widely used K-means algorithm in clustering. K-Means algorithm initializes the centroid randomly, which is where it could sometimes create less accurate clusters. K-Means++ overcomes initialization part to improve K-Means. K-Means++ algorithm takes an input k, which refers to the number of clusters that should be generated and n refers to set of objects [5].

K-Means++ clustering algorithm works as follows.

1. Select initial centroid X uniformly at random.
2. For each instance X we need to compute D(X) which is the distance between X and the nearest centroid that has already been chosen.
3. Choose next centroid using a weighted probability distribution which is proportional to D(X)2.
4. Repeat 2 and 3 until K centroids have been chosen.
5. Assign each record or instance point to a cluster centre which has least distance.
6. Calculate mean value of all points in the cluster.
7. Replace cluster centroid to this new mean value.
8. Repeat the steps from 6, until there are no more changes to centroids.

This algorithm can be implemented in map-reduce pattern as follows.

- *Map function:* The HDFS stores input data as sequence file of <key, value> pairs [6]. Every <key, value> pair represents a record. The map function splits the data across all mappers [7].
- *Reduce function:* After mapping, reducers are used for computing Step-2 of the algorithm. Reducer will also combine intermediate data of same mapper [8]. The intermediate data can be put in hdfs or stored locally [9]. New centres are generated which can be used for further iterations.

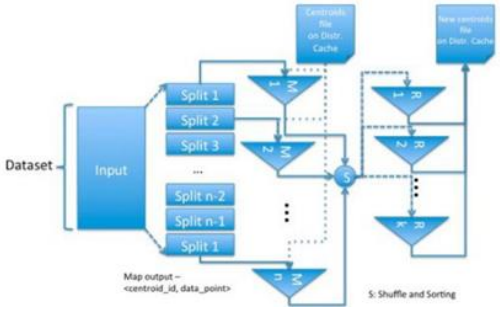

Fig. 4. K-Means++ Map Reduce [2]

4. Conclusion and future work

The Big Data generated today has demanded more computing, more storage resources as well as better way to process the data. Conventional methods were failing to cope up with such challenges because of which Cloud Computing and Hadoop MapReduce gained popularity [10].

The K-Means++ algorithm discussed in this survey is efficient for large set of data but it suffers with outlier issue. Future work could involve outlier detecting and removal algorithm merged with K-Means++ to give better and more accurate results.

References

[1] 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.
[2] 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.
[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] Borthakur. D “The Hadoop Distributed File System: Architecture and design”, 2007.
[5] Weihzong Zhao, Huifang Ma, Qing He, “Parallel K-Means clustering Based on MapReduce”, springer-verlag Berlin, Heidelberg, 2009.

[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.