International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 09 | Sep 2018 www.irjet.net p-ISSN: 2395-0072

Data Clustering Frame work using Hadoop

Malik Sadaf Allauddin[1], Abu Sufyan Malik[2]

[1]B.Tech, Dept. of IT, Pillai HOC College of Engineering and Technology, Maharashtra, India
[2]B.Tech, Dept. of Mechanical Engineering, MIST, Hyderabad, India

---

Abstract: - *Twitter has attracted millions of users to share and disseminate most up-to-date information, resulting in large volumes of data produced every day. However, many applications in Information Retrieval (IR) suffer severely from the noisy and short nature of tweets. A novel frame work for tweet clustering. By splitting tweets into meaningful cluster, the semantic or context information is well preserved and easily extracted by the downstream applications. For Clustering, distinct algorithms such as K-Means are suggested to cluster the data. There is a need for an algorithm that is able to cluster the data in a lesser amount of time, in case of data stream. Hence the need to use a parallel and distributed environment using Map Reduce framework, Likewise particle swarm optimization techniques are preferable for clustering problem, since it scales very well as data, dimensions increase. PSO algorithm for clustering twitter data using Hadoop map-reduce framework. The outcome illustrates that parallel PSO performs very well compared to K-Means algorithm.*

Key Words: Hadoop, Map Reduce, PSO, Twitter data, K-mean

1. INTRODUCTION

Clustering is performed on Twitter data using static algorithm as well as distributed algorithm using Hadoop framework. The problem behind sequential clustering algorithm is that they do not scale up with huge data size. And also many of them are considered to be poor in maintaining time complexities and memory space. Therefore in order to deal with very huge amount of data, one has to come up with parallelization technique so that time complexity and memory can be managed efficiently.

The analyzing sentiment of tweets comes under the domain of "Pattern Classification" and "Data Mining". Both of these terms are very closely related and intertwined, and they can be formally defined as the process of discovering "useful" patterns in large set of data, either automatically (unsupervised) or semi automatically (supervised). This would heavily rely on techniques of "Natural Language Processing" in extracting significant patterns and features from the large data set of tweets and on "Machine Learning" techniques for accurately classifying individual unlabeled data tweets) according to whichever pattern model best describes samples (them.

The features that can be used for modeling patterns and classification can be divided into two main groups: formal language based and informal blogging based. Language based features are those that deal with formal linguistics and include prior sentiment polarity of individual words and phrases, and parts of speech tagging of the sentence. Prior sentiment polarity means that some words and phrases have a natural innate tendency for expressing particular and specific sentiments in general. For example the word "excellent" has a strong positive connotation while the word "evil" possesses a strong negative connotation. So whenever a word with positive connotation is used in a sentence, chances are that the entire sentence would be expressing a positive sentiment. Parts of Speech tagging, on the other hand, are a syntactical approach to the problem. It means to automatically identify which part of speech each individual word of a sentence belongs to: noun, pronoun, adverb, adjective, verb, interjection, etc. Patterns can be extracted from analyzing the frequency distribution of these parts of speech (ether individually or collectively with some other part of speech) in a particular class of labeled tweets. Twitter based features are more informal and relate with how people express themselves on online social platforms and compress their sentiments in the limited space of 140 characters offered by twitter. They include twitter hash tags, retweets, word capitalization, word lengthening, question marks, and presence of URL in tweets, exclamation marks, internet emoticons and internet shorthand/slangs.

Classification techniques can also be divided into a two categories: Supervised vs. unsupervised and non-adaptive vs. adaptive techniques. Supervised approach is when we have pre-labeled data samples available and we use them to train our classifier. Training the classifier means to use the pre-labeled to extract features that best model the patterns and differences between each of the individual classes, and then classifying an unlabeled data sample according to whichever pattern best describes it.

1.1 Objectives of the Project

1. Analysis and clustering is the task to identify an e-text (text in the form of electronic data such as comments) to be positive opinion or negative opinion.

2. Instead of spending times in reading and figuring out the positivity or negativity of text we can use automated techniques for analysis and clustering.

2. Software Requirement

- Windows Operating System.
- Android SDK
- Eclipse (IDE)
- Java
- MySQL
- Hadoop 2.3.0

2.1 Hardware Requirement

- Processor: Pentium
- 4RAM: 4GB or more
- Hard disk: 16 GB or more
- Android Device

3. Module Description

1. Login
   In this module using username and password user login into system. In this login system authentication of user so only valid person login into the system.

2. Data Collection
   In this user select one book name from a books list and click on submit after submitting our system get reviews of this book using web mining technique. In web mining technique the system get data from another websites where reviews are present related with this book.

3.1 Clustering using PSO and K-means

The tweets retrieved using twitter API from twitter which can be cluster using K-means algorithm and PSO algorithm.

3.2 System Architecture

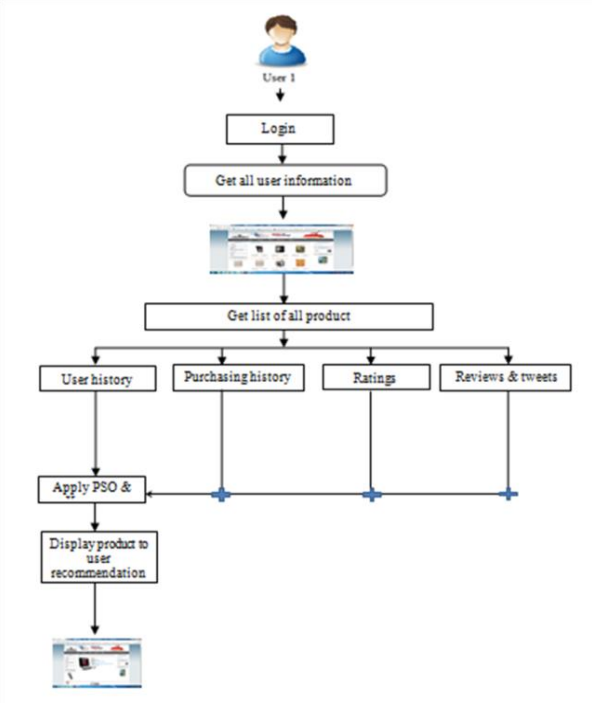


Fig –3.2: System Architecture

3.3Project Plan (Gantt chart)

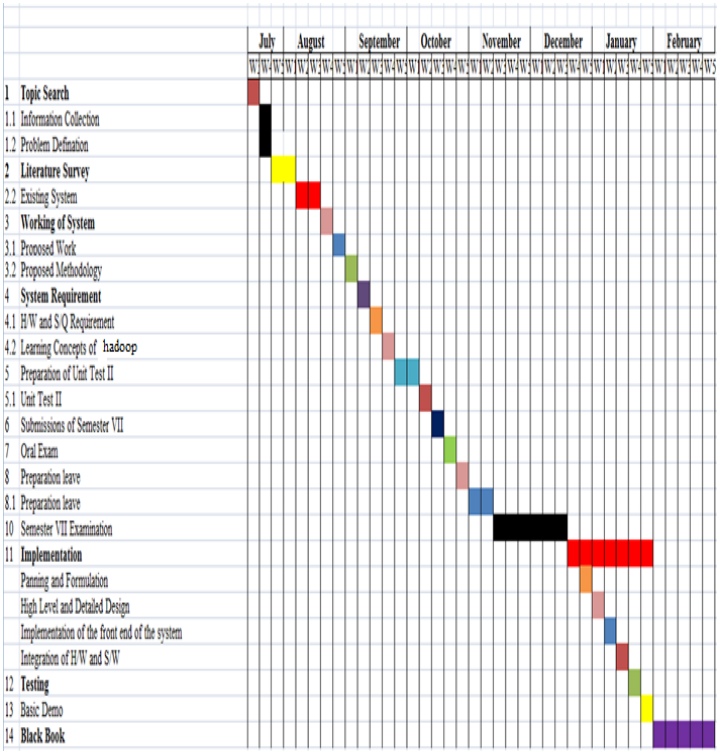


Fig -3.3: Gantt chart

4. Project Analysis

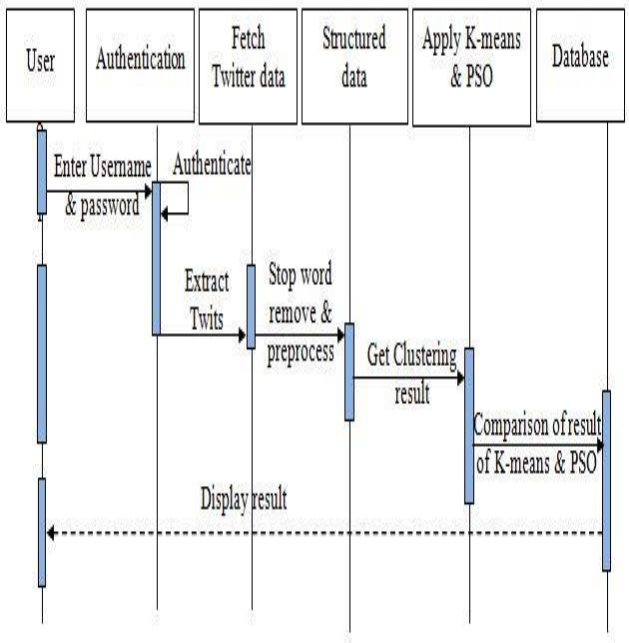


Fig -4: Sequence Diagram

5. Project Design

5.1 Design Model - Class Diagram

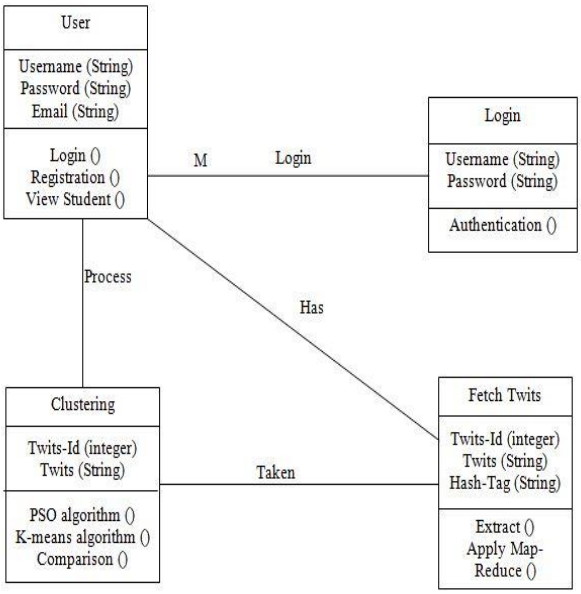


Fig -5.1: Class Diagram

5.2 DFD

A data flow diagram (DFD) is a graphical representation of the flow of data through an information system. A data flow diagram can also be used for the visualization of data processing (structured design). It is common practice for a designer to draw a context-level DFD first which shows the interaction between the system and outside entities. This context-level DFD is then exploded to show more detail of the system being modeled.

Symbols:

The four components of a data flow diagram (DFD) are:

_ External Entities/Terminators are outside of the system being modeled. Terminators represent where information comes from and where it goes. In designing a system, we have no idea about what these terminators do or how they do it.

_ Processes modify the inputs in the process of generating the outputs

_ Data Stores represent a place in the process where data comes to rest. A DFD does not say anything about the relative timing of the processes, so a data store might be a place to accumulate data over a year for the annual accounting process.

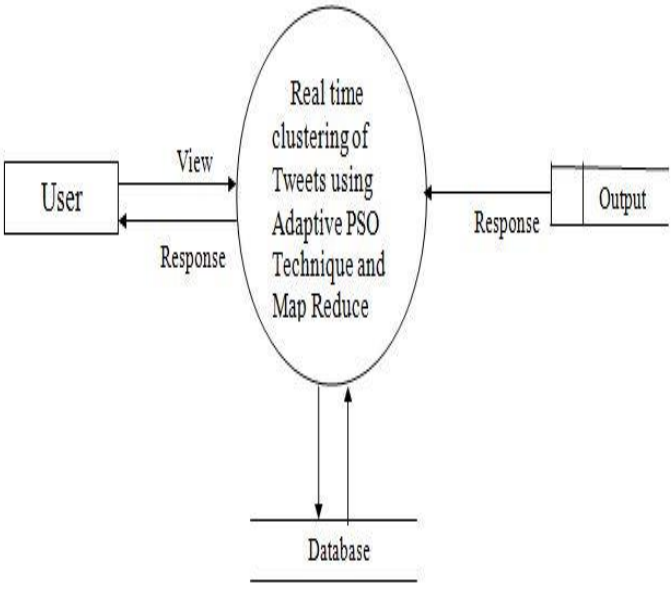


Fig -5.2.1: Level 0 DFD

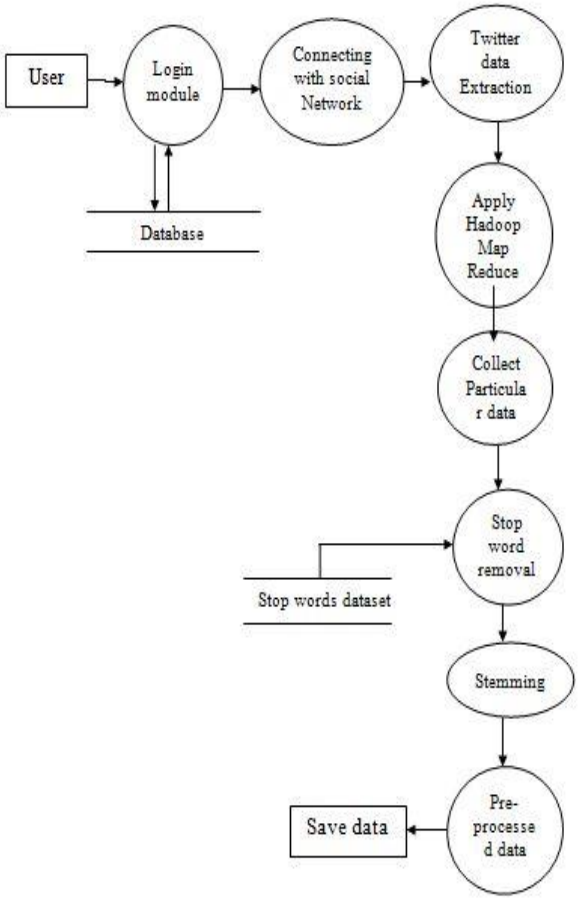


Fig -5.2.2: Level 1 DFD

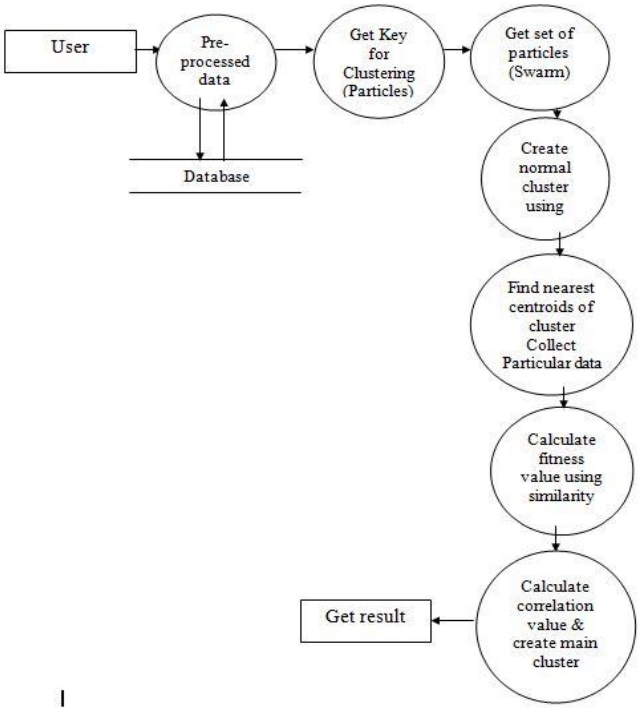


Fig -5.2.3: Level 2 DFD

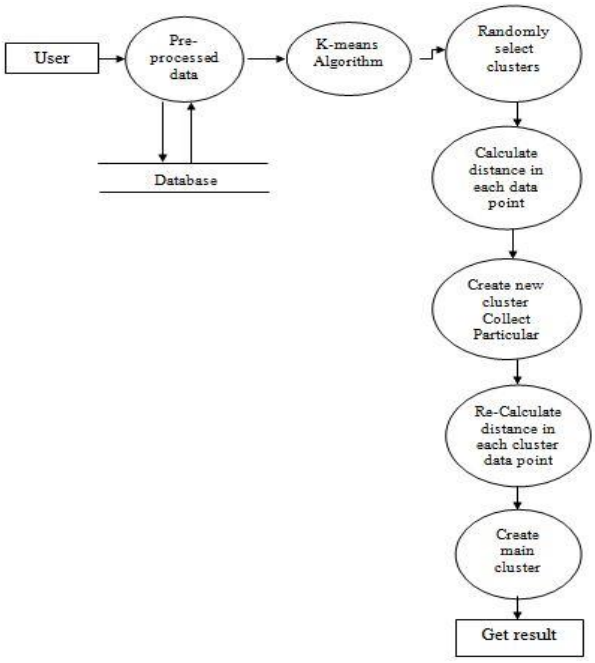
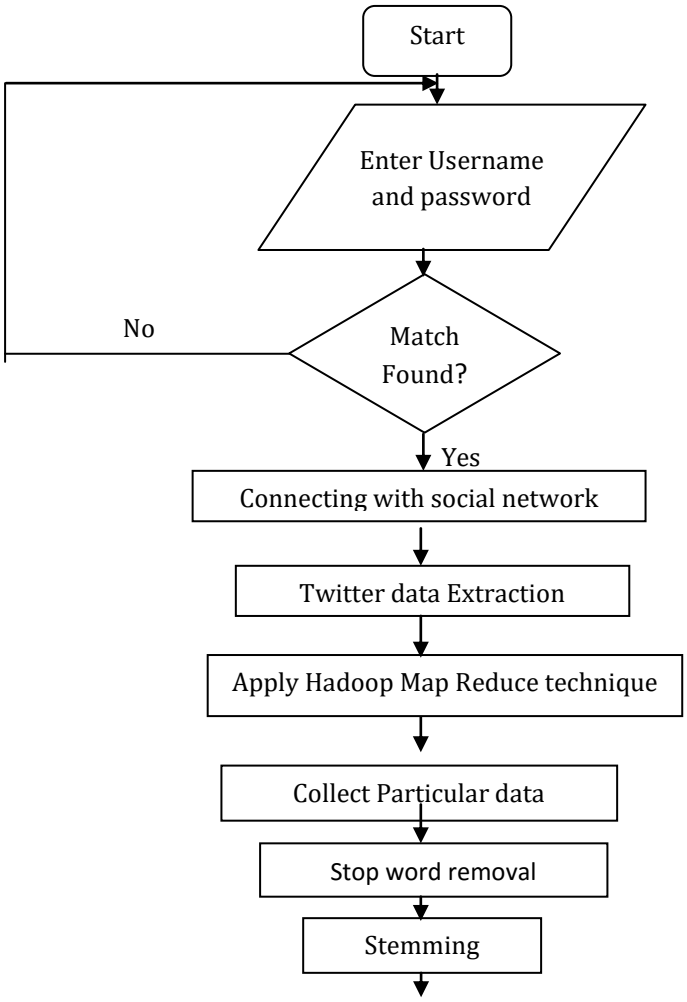
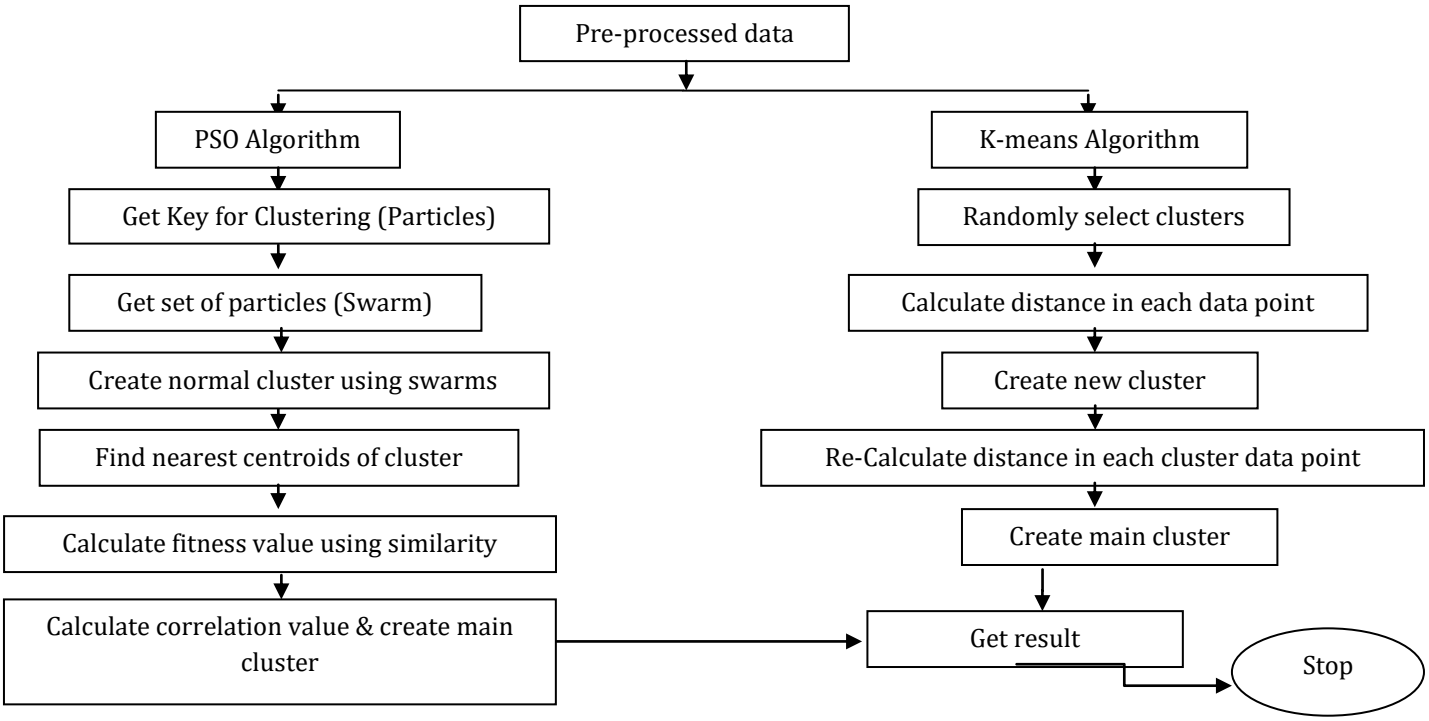


Fig -5.2.4: Level 3 DFD

5.3 Flowchart





6. IMPLEMENTATION METHODS

There are several methods for handling the implementation and the consequent conversion from the old to the new computerized system.

The most secure method for conversion from the old system to the new system is to run the old and new system in parallel. In this approach, a person may operate in the manual older processing system as well as start operating the new computerized system. This method offers high security, because even if there is a flaw in the computerized system, we can depend upon the manual system. However, the cost for maintaining two systems in parallel is very high. This outweighs its benefits.

Another commonly method is a direct cut over from the existing manual system to the computerized system. The change may be within a week or within a day. There are no parallel activities. However, there is no remedy in case of a problem. This strategy requires.

A working version of the system can also be implemented in one part of the organization and the personnel will be piloting the system and changes can be made as and when required. But this method is less preferable due to the loss of entirety of the system.

7. IMPLEMENTATION PLAN

The implementation plan includes a description of all the activities that must occur to implement the new system and to put it into operation. It identifies the personnel responsible for the activities and prepares a time chart for implementing the system. The implementation plan consists of the following steps.

1. List all files required for implementation.
2. Identify all data required to build new files during the implementation.
3. List all new documents and procedures that go into the new system.

The implementation plan should anticipate possible problems and must be able to deal with them. The usual problems may be missing documents; mixed data formats between current and files, errors in data translation, missing data etc.

8. ALGORITHM

Parallel PSO algorithm consists of 3 modules:

Module1: In this module, map reduce job is to update particle centroids. The algorithm for map-reduce 1st module is shown below.

Function Map (Key: particle's ID, Value Particle)

Initialization:

Particle's ID = Key,particle = Value

extractInf(VV, CV, PBC,GBC)

Generate two random numbers r1 and r2

For each Ci in CV do

For each j in Dimension do

New VVij= W * VVij + (r1 * c1)*(PBC - Cij) + (r2* c2)*(GBC - cij)

New cij = ci + new VVij

End for

Update(new VVi, new ci,paricle)

End for

Emit(particleID, particle)

End function

Function Reduce (Key: ParticleID, VaiList: Particle)

Foreach Value in VaList do

Emit(Key, Value)

End for

End function

Module2: For the newly updated swarm, new fitness values are calculated. Map function uses centroid vector for each particle and calculates distance between each records and centroids. It then returns minimum distance value with respective centroids ID. Map function forms new composite key by using particle ID and respective centroid ID that has the minimum distance value.

Module3: The goal of third module is to form new single swarm by merging results of first and second modules. New fitness value generated by taking average of all fitness values of centroids which were the results of the second module. Then the swarm is updated with newly generated fitness value. If new fitness value is less than the PBFV, then PBFV and its centroids are updated, Similarly if there is less fitness value than GBFV, then GBFV and centroids are also updated. Finally, new swarm with all new information is stored in DFS which will be the input for next iteration.

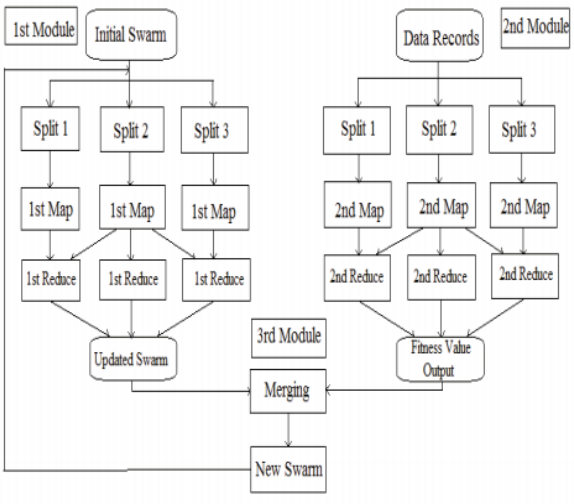
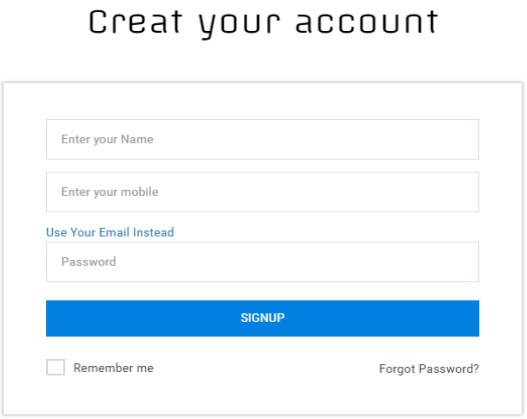
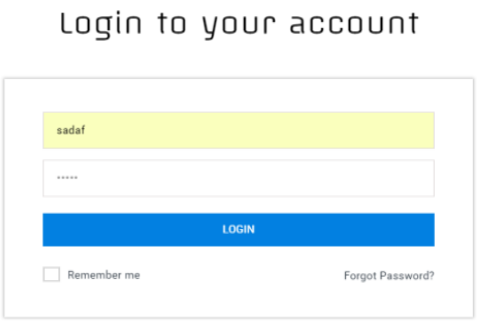
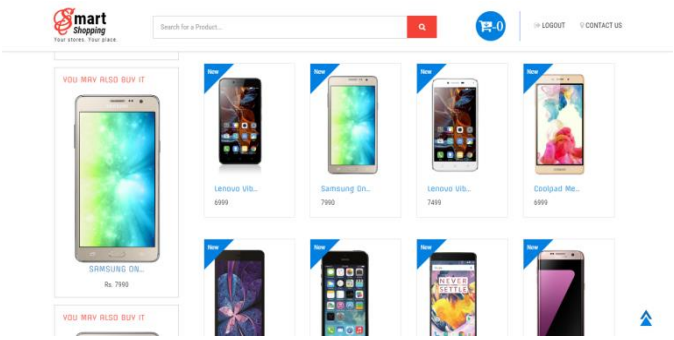


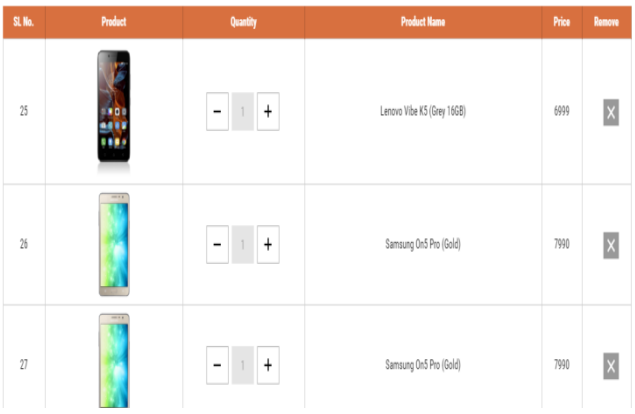
Fig -5.2.2: Map reduce model

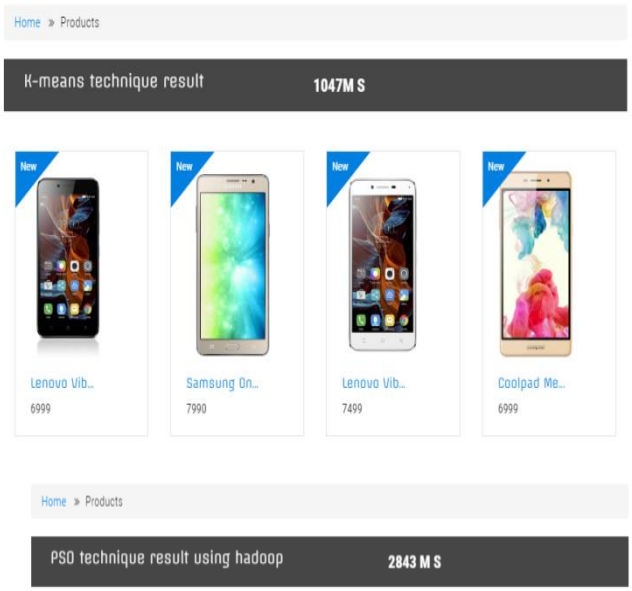
9. RESULTS











The execution time of the proposed parallel PSO Algorithm using different particle sizes.

From the given table we observe that:

i. The Execution time increases with the increase in the particle size.

ii. The Execution time decreases exponentially with the increase in the Number of nodes.

iii. The efficiency of the experiment is evident only up to three or four nodes.

10. FUTURE SCOPE

K-means clustering does not give better accuracy than PSO technique. However if we use K-means to generate initial seed points, the PSO algorithm is found to converge faster. Raising the volume of particles along with iterations increases the reliability, as it handles the problem space more efficiently.

11. CONCLUSION

Clustering results are tested on streaming twitter data using Adaptive Particle Swarm Optimization technique. This technique is usually adaptive because it does not entirely reinitialize particles while when fresh data falls by, improving the efficiency of execution. This method shows a trade-off between the cluster quality and number of particles used. K-means clustering does not give better accuracy than PSO technique. Raising the volume of particles along with iterations increases the reliability, as it handles the problem space more efficiently. K-Means clustering doesn't perform well with high dimensional data which is not the case with PSO Clustering algorithm. Hence it is ideal choice for text clustering.

REFERENCES

1. O. Shafique, "Recruitment in the 21st Century", International Journal of Contemporary Research in Business, Vol. 4, No. 2, 2012, pp. 887-901.

2. R. L. Bintliff, Crime proofing your business, New York: Mc- Graw Hill, Inc., 1994.

Ɓ. V. Brencic, and J. B. Norris, "Employers' online recruitment and screening practices", Economic Inquiry, Vol. 50, No. 1, 2012, pp. 94-111.

T
4. E. Parry, and S. Tyson, "An analysis of the use and success of online recruitment methods in the UK",
T Human Resource Management Journal, Vol. 18, No. 3, 2008, pp. 257-274.

5. A. Kar, and S. Bhattacharya, "E-recruitment and customer satisfaction: An empirical study in and around Kolkata", The Icfaian Journal of Management Research, Vol. 8, No. 2, 2009, pp. 34-54.

6. H. Sylva, and S. T. Mol, "E-recruitment: A study into applicant perceptions of an online application system", International Journal of Selection and Assessment, Vol. 17, No. 3.2009, pp. 311-323.