CHAPTER 4

ANALYSIS AND RESULTS

CHAPTER 4

ANALYSIS AND RESULTS

1.1. INTRODUCTION

The research work has been carried out with the acquired brain T2-weighted MR images. Parallel Genetic Algorithm is used with the average migration technique used for the received brain images segmentation in single Cluster Hadoop and Spark as well as Multi Cluster Hadoop and Spark in Amazon Web Services Cloud. Keeping the objectives in mind, the work has been done to achieve the primary objectives of the research. The research work is divided into two parts.

1. The first work focuses on enhancing the efficiency of the parallel genetic algorithm using average migration technique to segment the brain MR images for further analysis and diagnosis. It is explained in section 4.4.

2. Secondly, the research work proposes an architecture using Apache Hadoop and in the later stage with Apache Spark with two, three, four and five nodes in AWS. The carrying out of the proposed architecture "LISA" is explained in section 4.6.

It is well realized from the papers surveyed in the Literature Review section that there are very significant issues are evolving in PGAs. In terms of performance, coarse-grained PGAs are very complimenting but more complex in nature compared to the other segmentation techniques and serial GAs. There are different parameters influencing the migration of

individuals from any sub-population which is termed as deme to a different sub-population. For example, the topology defined to connect the demes for communication, the rate of migration of individuals among the demes and the interval in which the migration takes place are the basic as well as the important factors inducing the performance of PGAs. There is an evidence for the influence of migration rate in the PGAs found in the work given in the thesis of Grosso (1985). He has conducted an experiment on PGA with "delayed" migration scheme. In the delayed migration scheme, the author has made sure that the migration takes place till the population is near to converge. The process of brain MR images starts with image collection, storing, acquisition and preprocessing to prepare the images for the segmentation to get the optimized output using PGA.

1.2. BRAIN MRI SEGMENTATION USING PROPOSED MIGRATION TECHNIQUE

Segmentation of brain MR images is a cumbersome procedure which starts from acquisition and image preprocessing using various methods. Once the preprocessing is carried out, the image segmentation process takes place with the proposed average migration technique.

1.2.1. Image Acquisition

The first step of the image processing starts with the acquisition of images and preparing them for the further process. The collection of images and characteristics of the images have been discussed in chapter 3. As it is mentioned in the book of Daniel, Michael and McNamara (2012) the images received from online databases are processed in R for the conversion as a gray scale image of size 256*256 and then to a grainy

image in which Signal-to-Noise Ratio (SNR) is less. The converted 2D images are retrieved from HDFS and displayed in a two dimensional matrices where the elements of matrices are the pixels of the images. The size is important to decide the processing time and it takes the vital role for effective processing. In gray scale image, the values of pixels start from 0 to 255 where 0 signifies total black color and 255 signifies pure white color as Mustaqeem, Javed and Fatima (2016) mentioned in the image Journal. The values in between the range display the concentrations of gray color.

1.2.2. Images Storage into HDFS

Since the architecture uses Apache Hadoop to implement the algorithm, the Hadoop File System (HDFS) is used to store the images. HDFS may not be a right choice to store the small image files of size 100KB, which in turn will overburden the name node. For example, a full storage block size varies from 64 MB to 256 MB will be consumed by a small 10k .jpg file on HDFS. Meanwhile, it consumes capacity of namenode as a multi-terabyte file consumes which is unnecessary.

Storing image files into HDFS can be achieved by different ways. All the image files can be concatenated into a big chunk files and stored in HDFS. Then the offset of the big chunks is saved in a separate index. It makes easier the access of image later using location and offset of the image chunk from HDFS and load the chunk to memory for processing the file.

At the same time, there is a facility provided by Hadoop to read/write binary files. In the research, the MR images are considered as 256 X 256 gray scale values, the images can be stored into HDFS. There is a provision in Hadoop called SequenceFiles to complete this procedure of storing files.

Key/Value pairs in binary form are stored in a SequenceFile which is a flat file. There are classes in SequenceFile namely Writer, Reader and Sorter for writing, reading and sorting correspondingly. So, the file can be converted as a SequenceFile and stored it into the HDFS.

Since the size of a brain MR image is 80 to 100 megabytes in general, there is not a problem in storing the files in HDFS in either of the ways given. There are around 150 brain MRI files used for processing which are collected from cancerimagingarchive,net.

1.2.3. Image Pre-Processing

Image preprocessing plays a very important role in the field of image processing. As it was discussed before, the images scanned in different situations, patients and various external and internal parameters will have a direct impact on the segmentation results due to the artifacts present in the images. The primary data received from the authenticated hospitals would not have gone through preprocessing procedures where as some of the verified online databases provide preprocessed as well as the unprocessed brain MR images.

One of the best significant features for brain MRI segmentation is the level of intensity of the brain tissue. The variations in the brain tissue can help the most to segment the tissues to achieve the desired data. But there are possibilities for the intensity values to get corrupted due to the presence of MR image artifacts. So it is a mandatory process to take the MRI data through the preprocessing procedures to realize the best results in the segmentation using the proposed algorithm.

Usually, the image pre-processing prepares the image by eliminating noise and refining or aligning image quality for further processing intended. Since this research concentrates on segmentation, the common enhancement and noise reduction techniques were applied. The preprocessing techniques in this context have been implemented by Shen, Sandham and Granat (2003). The image enhancement helps in reducing the salt and pepper effect from the image which might cause errors in future processing. These are various preprocessing methods and proposed by different authors which include the functions such as de-noising, skull-stripping, intensity normalization, etc. The following sections will define the different steps followed to pre-process the MR images for the segmentation.

1.2.4. Noise Removal (de-noising)

Noise removal denoted as de-noising in image preprocessing is a standard procedure for MRI. Since noise present in MR images will mislead the results of segmenting tumor from other normal tissues in the brain, it is mandatory to decrease the noise to the maximum and to augment contrast between regions.

To remove the noise in the given original image, any filtering tehnique can be used. Since the research is focusing on tumor detection, the sharpness of the edges must be measured as a pivotal point and it must be preserved. The sharpness of the edges can be realized by the abrupt modification of intensity in the gray values of pixels.

There are different de-noising methods like Anisotropic Diffusion Filtering (ADF), wavelets, Non-Local Means (NLM), and Independent Component

Analysis (ICA) used for MR images preprocessing. Among these methods, ADF is a famous and commonly used method for the de-noising of brain tumor MR images.

Perona and Malik (1987) presented anisotropic diffusion also termed as Perona-Malik diffusion as a multi-scale technique to reduce noise in images without eliminating substantial parts of the image intended for processing normally such as lines, edges and other details which are significant for the image interpretation of the to perceive edges. The primary notion of the algorithm is to provide smoothing within the boundary and continuous regions of the image but to skip smoothing the regions crossways boundaries in the image.

In ADF method, the edges already executed are considered as regions with diffusion coefficients of lower value and the filter works on the basis of a constrained differential diffusion equation. Following is the diffusion equation:

$$I_t = \mathbf{div}(c(x,y,t)\nabla I) = c(x,y,t)\Delta I + \nabla c \cdot \nabla I \qquad (1)$$

div - divergence

∇ - gradient,

Δ - laplacian

I - intensity image.

t - diffusion time

c(*x; y; t*) - a scalar field which controls the power of diffusion

ΔI - monotonically decreasing relative function to the initial degree of the gradient.

The initial value of function $c(x; y; 0)$ is near to zero at locations with large gradients and in such locations, boundaries are supposed to occur. The value of function $c(x; y; 0)$ is found maximum at locations where the gradients are small.

The following equations (2) and (3) are proposed by Perona and Malik to evaluate the values of $c(x; y; t)$:

$$g(\nabla I) = \exp\left(-(\|\nabla I\|/K)^2\right) \quad (2)$$

or

$$g(\nabla I) = \left(1 + (\|\nabla I\|/K)^2\right)^{-1} \quad (3)$$

In the equation, K is a scalar parameter which controls the edge enhancement threshold.

It is evident from the study that there are adverse effects in the brain tumor segmentation due to noise present in minimum in the image even after applying noise removal procedure.

1.2.5. Skull Striping

Skull stripping is one of the best preprocessing methods among the various methods adopted, mostly for the segmentation of brain MR image. The cortex is the outermost layer of the brain composed of folded grey matter which can be pictured as a discrete dark ring appears in the brain MR images, encircling the tissues. Skull stripping is a procedure to retain the

brain soft tissues by removing skull, scalp, and meninges which are noncerebral tissue regions not required for image processing.

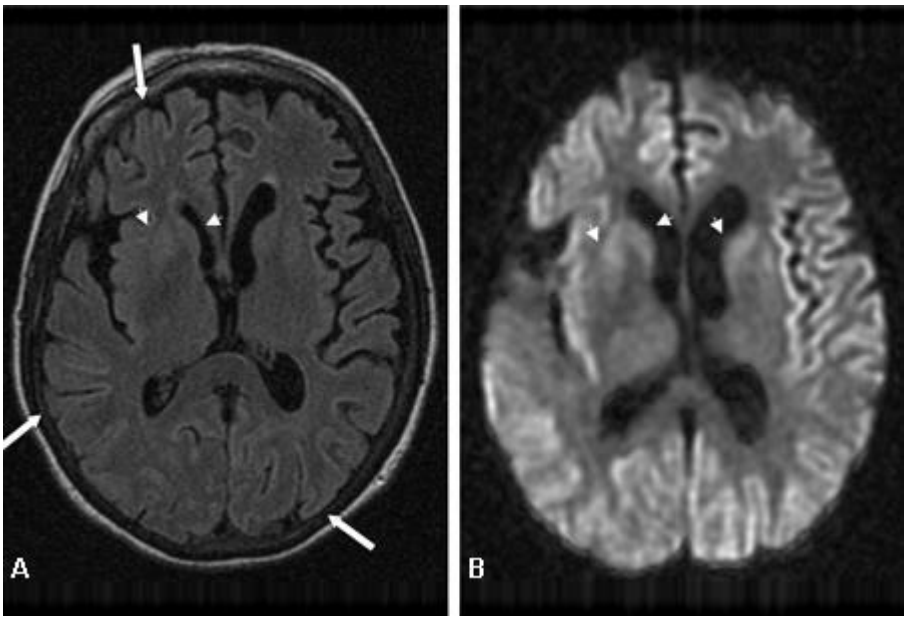

Figure 4.1 Cerebral cortex gyral hyperintensities showed in MRI.

The first step of the skull stripping method starts with the conversion of the given MRI brain image into gray scale image. The morphological operation implemented by Soumya (2011) is accomplished in the converted gray scale image by using contrast adjustment.

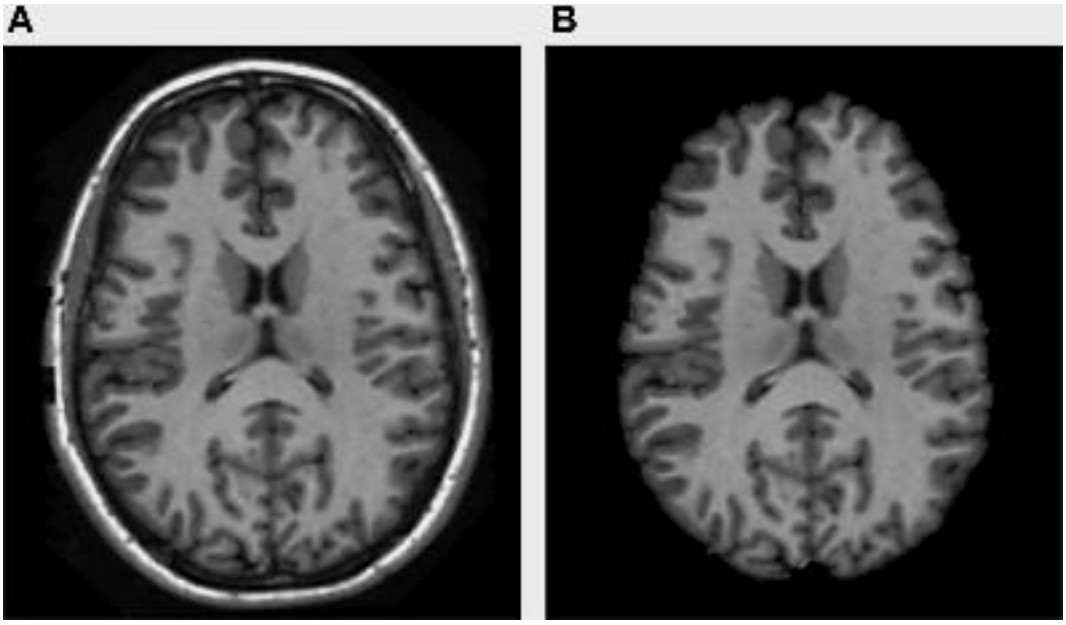

Figure 4.2 (A) Original MRI scan (B) After skull-stripping Procedure.

There is an amazing enhancement in the accuracy of the procedures to detect a tumor, plan for surgery, and reconstruct cortical surface and brain morphometry because of the accuracy in implementing skull stripping procedure which is studied by Demir and Sayar (2014). Due to its efficient involvement in MRI preprocessing, skull stripping is measured as an inevitable step in the process of brain tumor segmentation as studied by Chan et al. (2015). Moreover skull region removal from the brain MRI increases the probability of classification of pathologies.

There is a challenge while adopting skull stripping procedure on the MR images. The homogeneity in intensities of non-cerebral and the intracranial tissues may cause the removal of the necessary information during segmentation. It may lead to a negative impact on the selection of seed point followed by the inaccuracy in end results. The complex structure of the human brain, inconsistency in the factors of MRI machines, and the

patients' prevailing and varying conditions are also influencing the results of skull stripping procedure. Region based binary mask extraction is used by Selvi and Duraiswamy (2013) to strip of the brain cortex in the gray scale image.

After the noise removal, the image is taken for sharpening of the edges. The earlier studies show that the median filter is most suitable for noise removal in brain MR images since it is best in working with the salt and pepper noise without interfering too much with the color contrast. It has been studied from the publication of authors Jian and Bovik (2002).

1.2.6. Image Sharpening

High pass filters are generally used in image sharpening. It is vital that the edges must be kept as they highlight the tumor in the processed image. To improve the boundaries of the objects, the Gaussian filter, a high pass filter is used. The research uses the Gaussian filter for image sharpening and this filter is also used broadly.

In image sharpening, the main objective is to enhance details which are blurred due to image artifacts or to high point the important details in the image. The objects to be extracted from an image for future analysis are improved with the usage of image sharpening and this indicates a spatial filter shape where high positive object takes the center place.

The image sharpening matrix of a simple spatial is as follows.

Table 4.1 Spatial filter using image sharpening

-1/9	-1/9	-1/9
-1/9	8/9	-1/9
-1/9	-1/9	-1/9

The weight assigned to each high positive object is summed up and the total value is zero. In Fourier expansion, this zero DC value is the coefficient of the zero frequency term. The value of an offset is to be kept in the 0....255 range for display purposes.

a) High Pass Filtering

In high pass filtering, a low pass image is subtracted from the original image. It is denoted as follows.

High Pass Image = Original Image - Low Pass Image

In many cases, some low pass objects are remembered even though a high pass image is compulsory, to support in the analysis of the image. In such cases, if amplification factor (AF) is multiplied with the original image before deducting the low pass image, the exercised filter is labeled as high boost filter or high frequency emphasis filter. The following equations represent the calculations for applying the filters.

High Pass Image = AF. Original Image – Low Pass Image

= (AF- 1) . (Original Image) + Original Image – Low Pass Image

= (AF-1) . Original Image + High Pass Image

In the above equations,

If AF = 1, it is termed as a simple high pass filter.

If AF > 1, fragment of the actual input image is maintained as it is in the output.

A simple filter for high boost filtering is given by

Table 4.2 Simple high pass filter

-1/9	-1/9	-1/9
-1/9	ω/9	-1/9
-1/9	-1/9	-1/9

Where ω = 9A-1.

b) Median Filter

It is necessary to perform noise reduction of a high degree on an image in medical image processing before high-level processing steps are performed on it. Median Filter is greatly used in noise removal process in high frequency MR images without altering the edges. This procedure estimates the median of the adjacent pixels to decide the new demonized value of the pixel. A median is estimated by organizing all pixel values by means of their size and the new median value for the pixel is assigned.

For each pixel, the pixels in the adjacent window are extracted, and the intensity values of the pixel are organised in ascending order. Now the median value for the new window is to be formulated. The median value replaces the intensity value of the centre pixel. It has to be continued for all

the pixels in the MR image to smoothen the edges and the final image is enhanced with high resolution.

Image segmentation is the next and important step in image processing after image enhancement. The improved and enhanced image can produce better results while processing for edge detection and quality improvement of the image during Image Segmentation.

1.3. IMAGE FEATURES

The features of an image reveal unique appearances of a specific image to be segmented. The features are identified by the mathematical calculations which comprise of shape descriptors and quantitative visual appearance. These features enhance the discrimination of structures and background required for further analysis. The success of any segmentation procedure primarily depends on the selection of precise and suitable features. It has been well discussed in the work of Ivana Despotović, Bart, and Wilfried (2015).

It is common that the neighbouring pixels will have almost the same feature since the pixels in an image are highly associated with each other. Due to this reason, the spatial connectivity of the adjacent pixels can be considered as one of the most important features for segmenting the desired data from MR image.

Feature extraction and classification in MRI brain images are achieved in common through statistical methods. The features extracted by the statistical methods represent the image texture as a vector in a multidimensional space. The first and second order values of gray level

intensities in the image forms the statistical features set. The mean, median, standard deviation and the intensity derived from the pixel values are considered as first order statistical features. At the same time, the first order features alone are not enough since they do not provide a spatial distribution of the pixel values in an image. To achieve the best results, the second order features are also combined with the first order feature to describe the image texture in depth. By means of gray level co-occurrence matrix, second order features are calculated and they are used to define the image texture. In the publication of Haralick R. M., Shanmugam K., and Dinstein (1973) first and second order features are termed as appearance features since they are the representors of optical appearance of an entity.

Following are the features defined for the segmentation.

Table 4.3 Image Features

Base	**Features**
Texture	Contrast, Correlation, cluster shade, Entropy, Energy, Homogeneity, sum of square variance
Intensity	Mean, Variance, Standard Variance, Median Intensity, Skewness, and Kurtosis
Shape	circularity, irregularity, Area, Perimeter, Shape Index

In segmentation of medical images, using probabilistic prior shape models can also add value to the performance boost and it is being used widely. We could see it in the work of Yan et al. (2010). An average shape and deviations occur in the object of interest are quantified by probabilistic

prior shape models adopted in Tao, Prince, and Davatzikos (2002) and they are assessed from a population of images of the object associated together.

4.4 PARALLEL GENETIC ALGORITHM FOR IMAGE SEGMENTATION

One of the main objectives of this research concentrates on using Parallel Genetic Algorithm, to segment the given brain images into a set of semantically meaningful, homogeneous, and non-overlapping regions of similar attributes such as intensity, depth, color, or texture to identify the tumor in the brain. First order features like individual pixel/voxel intensities play a vital role in brain image segmentation to find the tumors.

Parallel Genetic Algorithm performs well with respect to efficiency especially the search space is complex and large where accuracy is more expected. PGA works more natural by integrating the migration operation of best fit individuals from different subpopulations generated in the various worker nodes.

The following is the basic flow of the parallel genetic algorithm.

1. Definition of genetic operators.
2. Random generation of a population of candidate solutions.
3. Partition the populations into several subpopulations.
    a) Apply the average migration strategy for individuals to flow amongst the subpopulations.
4. Execute the following steps (a) and (b) for each subpopulation.

a) Depend on the selected genetic operators selection, crossover, and mutation, execute self-evolution.

b) Based on the average migration technique, the best individuals are sent to the subpopulations, where the best ones replace the worst ones of the subpopulation.

5. Check whether the stopping conditions are satisfied. If fulfilled, break the iteration, else go to Step 3.

Performance improvement can be achieved by using the fitness valuation of parallel GA with respect to the data size and the number of clusters used. The efficiency of parallel GA can be enhanced with the fitness evaluation, migration rates, communication topology, or deme size. The time needed to reach the solution and the accuracy of the solution depends on the population size which is one of the important parameters in the proposed PGA. The research takes migration scheme as a vital parameter that decides which individual can migrate from one deme to another and which individuals are replaced.

4.4.1 Topology Used

It is a common practice to adopt static topologies for communication which will not be changed throughout the execution of the multiple population PGAs. The topologies are chosen in favour of the native topologies accessible for the investigators. Tanese (1987) has implemented hypercube topology which is commonly used by the researchers for the execution of PGA with fixed migration rate.

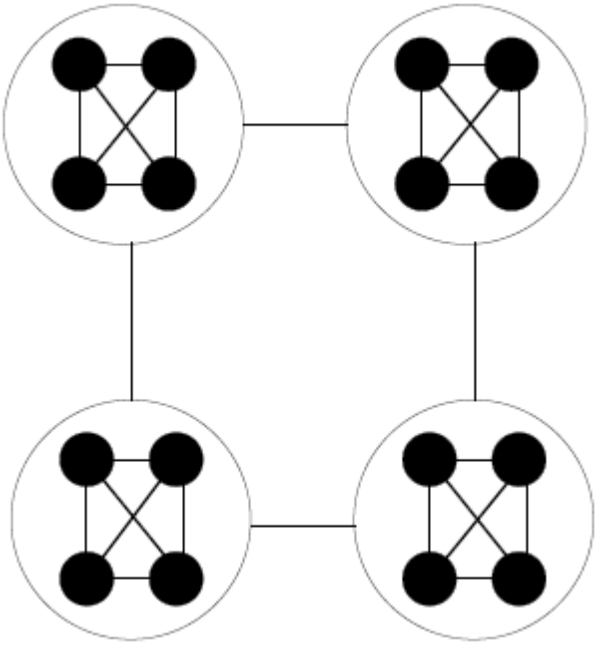

Figure 4.3 Fully Connected Topology.

4.4.2 Proposed Average Migration Technique

The island model and the stepping stone model are the most popular methods used in migration schemes. With respect to island model, a best fit individual of any deme can migrate to any other deme as happens in real world. But the cost of the communication with respect to time will be more since the communication may happen between any demes. The other most popular method called stepping stone model restricts the best fit individual to communicate with the two neighborhood demes to reduce the communication overhead. But there is a need to compromise the real world experience of extended communication.

We have taken the algorithm proposed by Grosso (1985) where the migration rate was fixed and the population with the partitions of five demes. It was observed that demes still worked autonomously and reconnoitered diverse regions of the search space with a low migration rate and it was like island model. But with an intermediate rate of migration the solutions were similar to those found in the panmictic population.

There is another methodology proposed by Tanese (1989) where the migration occurred in a fixed interval irrespective of the number of iterations or the crossed generations. The migrants among best fitness individuals were selected probabilistically and sent them to the receiving deme to replace the worst in them. The research has been carried out in three different conditions where the third one with the effect of the exchange frequency on the search. Results showed that too frequent or too infrequent migration tainted the performance of the algorithm.

By taking into consideration the advantages and disadvantages of these three models, the proposed model shows improvement in reduction of time by reducing the communication length, meanwhile retains the real world scenario. In the proposed methodology, the neighborhood space is extended to the length of eight populations and each subpopulation of size 16. There is a probability of Crossover 0.9 between any subpopulations. The fitness of each individual is calculated with the help of objective function which is exercised from the given features. The migration rate is denoted by the percentage of the set of individuals which can migrate between particular times.

Migrants were selected with the selection function and passed to one neighbour using a hybrid topology. Migration rate was initialized with 8 which is the half of sub population size 16.

The algorithm for the Average Migration Technique is given as follows.

```
initialize  max_gen 1024
initialize  pop_size 8
initialize  sub_pop_size 16
initialize  num_of_neighbours 3
initialize  migration_rate  8
initialize  num_of_ migrants  2
begin
    for i=0  to sub_pop_size
    //parallel  execution over sub-populations
    { initialize  function call
    // sequential process
    }
do{
        calc_migration_rate()
        {
        if (num_of_generations >8)
```

```
        migration_rate = (2^pop_size)/sub_pop_size + (1024-
        num_of_generations)
        }
      for j=0 to migration_rate
      {
      selection(..) //sequentially selected based on the fitness value
      crossover(..);
      mutation(..);
      }
   for j=0 to num_of_migrants
    emigrant[j] = select_emigrant(..);
//sequentially happens in the same node
      for j=0 to num_of_neighbours
      {
      send_migrants(..);
      // parallel execution to send & receive migrants
      receive_migrants(..);}
}while (generations <= max_gen);
}
end
```

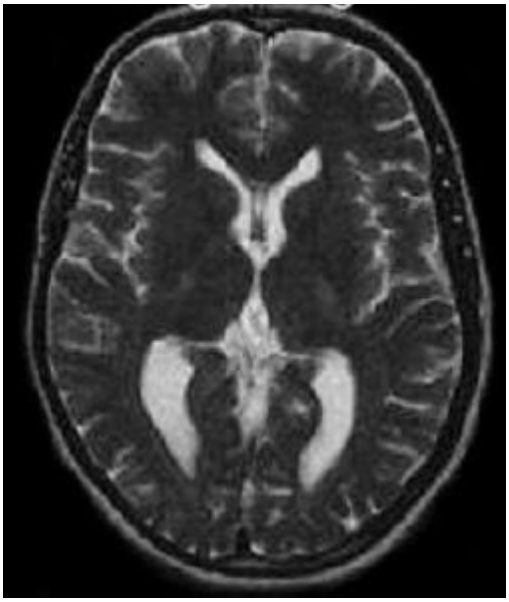

Figure 4.4 Normal - T2W MRI Brain in axial plane

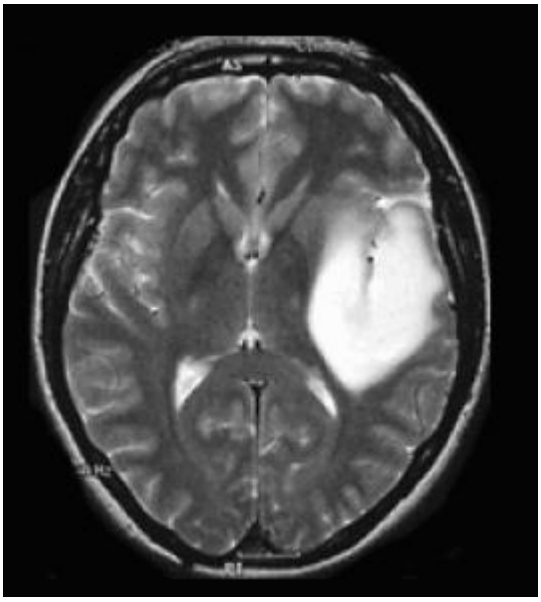

Figure 4.5 Benign Tumor - T2W MRI Brain in axial plane

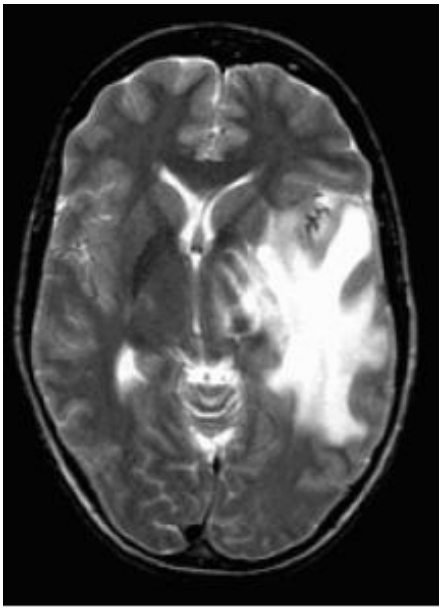

Figure 4.6 Malignant Tumor - T2W MRI Brain in axial plane

4.4.3 Segmented Images Output

There are 250 brain images grouped for research and the research has been tested with 150 brain MR images where 40 are normal brain MRIs and 40 brain MRIs are benign which cannot be considered as a tumor. The remaining 70 brain MRIs are with malignant tumors and known as low grade glioma.

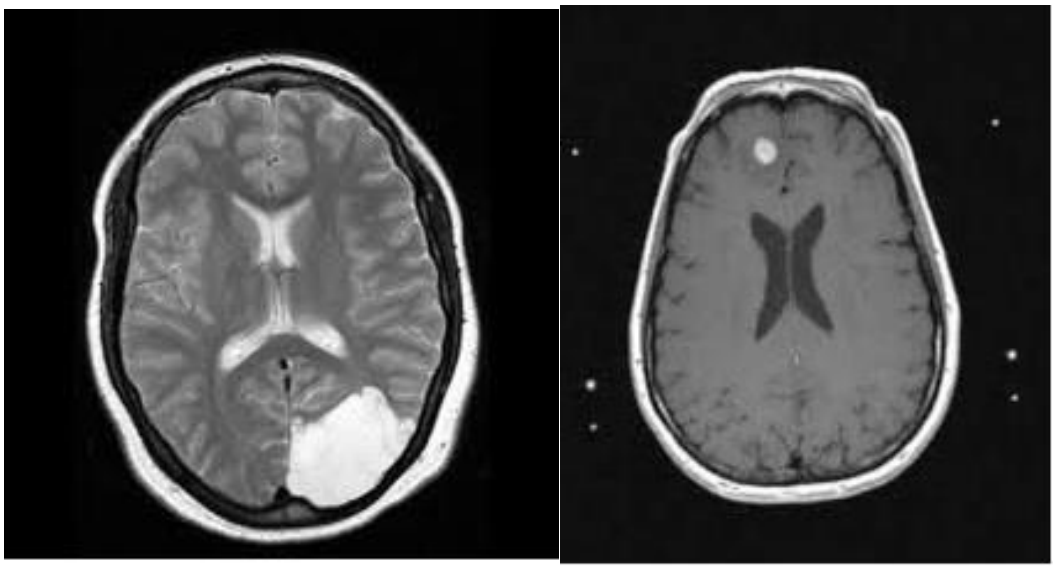

Figure 4.7 Segmented T2-weighted brain image with tumor

Even though there are various algorithms, SVM is used to test the accuracy with the intention that it could be beneficial while using high dimensional spaces. The images used for testing are T2-weighted and the size of each image is 256*256. These images were acquired at more than a few positions of the trans-axial plane. These images were processed using the PGA with the proposed migration technique using Apache Hadoop programming environment.

4.5 SVM FOR IMAGE CLASSIFICATION

Support Vector Machine (SVM) is used for image classification based on the features attributed. SVM was first developed by developed as an extension of the Generalized Portrait algorithm Vapnik and Lerner (1963) works on a learning algorithm which is constructed on the statistical theory.

SVM is a supervised classifier and uses the theory of statistical learning to reduce the structural risk.

The functions of SVM kernel is to increase the margin between the classes to the maximum and lessen the true costs by controlling the classification capacity and empirical risk. SVM tries to get the optimal hyper-plane which separates the members of a given class from the non-members using the multi-dimension space consists of features.

There are two steps in SVM classification:

1. Training Data
2. Testing Data

SVM prepares itself to proceed the classification by using the learning algorithm which takes a set of features as input. SVM finds the best margins amongst two classes to differentiate them during training of data. The different features described for classification are labelled in relation to class associative with a specific class. Because of the idea of hyper planes, SVM removes the problem of local minima and selection of neurons for every problem uniquely.

To enable the classification, the SVM algorithm uses the feature subset predefined in the research. There are three classes of images namely normal, benign and malignant. SVM applies the classification procedure using training set and test set data to evaluate the accuracy of segmentation by the proposed algorithm. In every processed image, individual subject is denoted by a vector.

There are four basic common kernels found in SVM as follows.

Table 4.4 SVM Kernels

Kernel Type	**Function**
Linear	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
Polynomial	$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)^d,\ \gamma > 0$
Radial Basis Function (RBF)	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \lVert \mathbf{x}_i - \mathbf{x}_j \rVert^2),\ \gamma > 0$
Sigmoid	$\tanh(\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)$

γ, r and d - the kernel parameters.

x_i are denoted as training vectors.

ϕ – Function that maps training vectors x_i to higher dimensional space.

$C > 0$ – Penalty parameter of the error term.

$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ - Kernel function.

SVM attempts to find all possible hyper planes to maximize the margin to reduce the training error as much as possible in the case of linearly separable data. The hyper plane separates the data used for training from their nearby points with the maximum distance. The following three major factors will reveal the efficiency of the algorithm.

- Sensitivity – Probability for a test is to be positive when a given input image has a tumor.

- Specificity – Probability for a test is to be negative when a given input image is normal.
- Accuracy - Probability for a test is accomplished appropriately.

SCM does not work with reducing the objective function dependant on input training data. But SVM has the advantage of minimizing the degree of error erupted while working with the test data by the learning machine. Because of this feature, SVM has the calibre to achieve results well when it works with test data. SVM concentrates more on a difficult task of training set data to get the best classification results. The training data falling on the "borderline" are called support vectors.

As it is mentioned earlier, three important factors sensitivity, specificity, and Accuracy are considered for efficiency evaluation of the proposed PGA. There are four cases namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) used to measure the above three factors. TP signifies correctly categorized positive cases, TN represents correctly categorized negative cases, FP represents incorrectly categorized positive cases and FN signifies incorrectly categorized negative cases.

The total number of brain MR images is 150 used for training in this research. In the set of images, there are 40 normal images, 40 images with a benign tumor and 70 images with a malignant tumor. Table 1 shows the classification rates for performing the proposed hybrid approach and comparison with different methods. The proposed PGA with the set values has shown the improvement by at least 0.67% in accuracy.

Table 4.5 Classification of percentage rates for the Proposed Method

	Island Model - Migration occurred among the adjacent demes alone	**With a fixed migration rate exchanged individuals with all the others**	**Migration occurred at fixed intervals between processors**	**Average Migration between demes and with processors**
Test Results				
Positive				
True positive	66	67	68	68
False negative	4	3	2	2
Negative				
False positive	1	1	1	1
True negative	79	79	79	79
Output:				
Sensitivity	94.29%	95.71%	97.14%	98.57%
Specificity	98.75%	98.75%	98.75%	98.75%
Accuracy	96.67%	97.33%	98.00%	98.67%
Positive predictive value	98.51%	98.53%	98.55%	98.57%
Negative predictive value	95.18%	96.34%	97.53%	98.75%

In medical image segmentation especially brain tumor segmentation, it is difficult to draw a gold standard. It would be advisable to take contours from 4 to 8 experts, from a clinical point of view. This is the reason, it is necessary to bring the expert system in segmentation to compare the automated contours with the ground truth contours. The second part of the architecture can help in creating an expert system to compare the results with the ground truth.

4.6. PROPOSED ARCHITECTURE USING HADOOP/SPARK

The second major objective of the research to suggest a "Next Generation Healthcare System" which is accomplished by using the Apache Hadoop and Spark in AWS. Easy and fast comparison of the features learned from the given segmented image with the existing huge volume of already processed data of segmented MR images to give an in-depth prediction to the doctor. The above methodology which was implemented in single cluster Apache Hadoop is implemented in the AWS with multiple clusters Hadoop and later with Spark.

4.6.1. Apache Hadoop2.4.0 in AWS Cloud

Distributed Processing of huge sets of data across the clusters of computers is well achieved with Apache Hadoop software library framework which uses MapReduce model. Apache Hadoop has the caliber to step up from a single server to thousands of machines. Each machine is considered as an instance and delivers local computation and storage. The Hadoop library is capable of diagnosing and handling failures at the application layer. Pierfrancesco, Mariano, and Paolo (2015) has explained this mechanism of

Hadoop which provides high-availability. The server is denoted by Namenode and the client machines are illustrated as Datanodes.

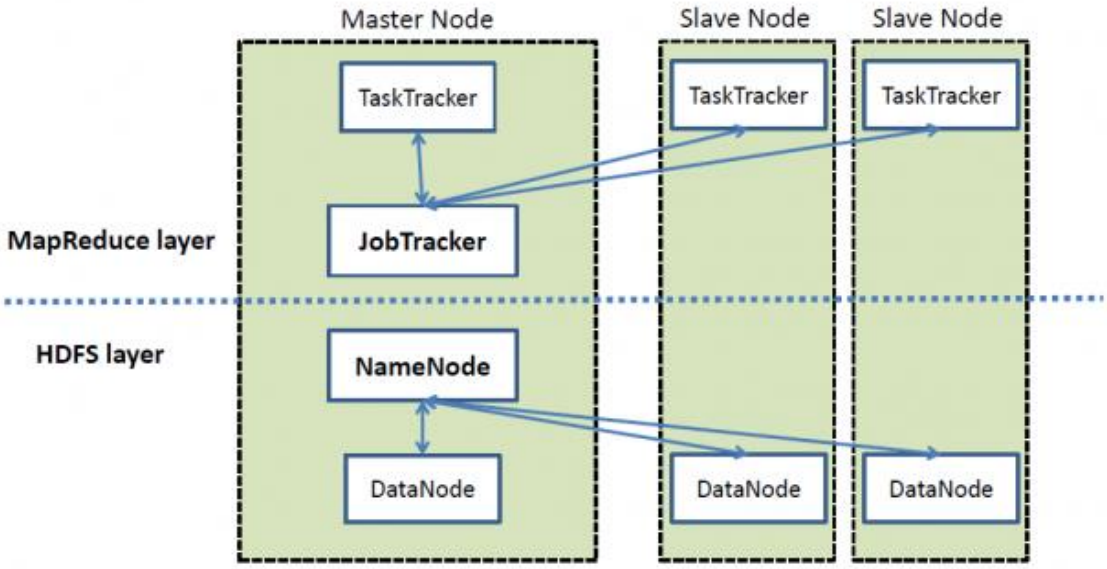

Figure 4.8 High Level Architecture of Hadoop (Sachin, 2014)

The following are some of the basic modules present in Hadoop framework.

- **Hadoop Common**: It contains the utilities commonly used to support further Hadoop modules.
- **Hadoop Distributed File System (HDFS)**: It is a distributed file system based on Java file system to store data and it gives highly scalable and reliable storage of data.
- **Hadoop YARN**: It is a framework to support resource management and job scheduling and monitoring by breaking up their functionalities into isolated daemons.

- **Hadoop MapReduce**: It is a system to enable the parallel processing of data sets.

In the first level, Hadoop Distributed File System (HDFS) is used as a prime storage system to store the brain MRIs of around 10 GB. HDFS supports reliable and rapid computation by creating multiple replicas of data blocks and allocates them on compute nodes all over a cluster.

4.6.2. Spark1.4.0 in AWS cloud

The proposed methodology is also implemented in Apache Spark to enable the process much faster since the volume of medical image data is huge and later the process can be on voxels.

Spark is a framework which facilitates cluster computing in a fast manner. Since it follows the map reduce mechanism, it is well compatible with Apache Hadoop. Any format of data working with Hadoop can be used in Spark and it can manipulate data stored in a storage system defined for Hadoop. One of the important characteristics of Spark is in-memory computing which improves the efficiency by means of execution time. Since Spark inherits the features of Scala, Java and Python, the efficiency is increased extraordinarily. Spark has been tailored as a one-stop solution for flexible and ultramodern data analytics by the continuous and numerous enhancements such as Spark streaming. Spark streaming supports the execution of the sequence of batch jobs which are very small and deterministic.

Since Resilient Distributed Datasets are kept in memory Spark Streaming has impressive performance. Spark can process data from a range of data

sources including HDFS, NoSQL databases such as Hbase and relational data stores such as hive. It is referred to the work of Harnie et al. (2015) that Spark can process data is in memory as well as data on disk and take advantage of data locality. As Spark has the interfaces to Java, the proposed system uses Java.

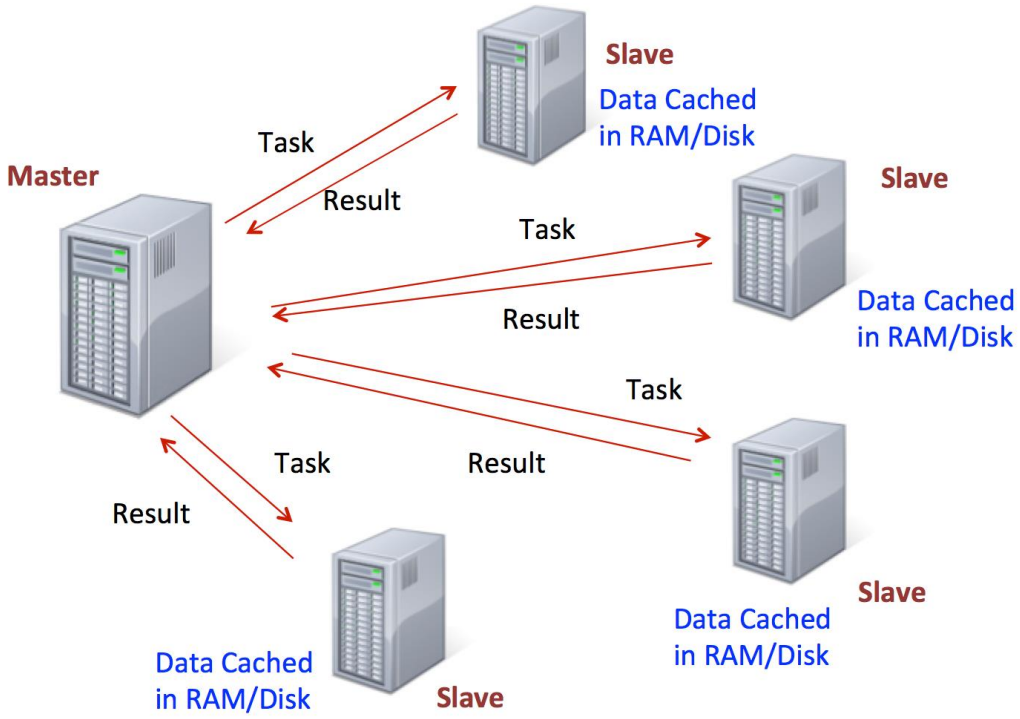

Figure 4.9 Job Execution model of Spark (Ashwini, 2015)

4.7. PERFORMANCE EVALUATION

The following architecture was implemented with single cluster Hadoop, Multi Cluster Hadoop, and the recent in-built MapReduce Spark. The performance of the PGA was tested with approximately 2 GB, 4 GB, 6 GB, 8 GB and 10 GB brain MR images stored in HDFS.

The algorithm was tested with 8 GB RAM, 1 TB HDD, Intel Core i7 CPU with 2.50 GHz speed, Ubuntu 14.0 and Hadoop2.4.0 configuration for the single cluster Hadoop. The same configuration was maintained for the multi cluster Hadoop with five DataNodes and one NameNode.

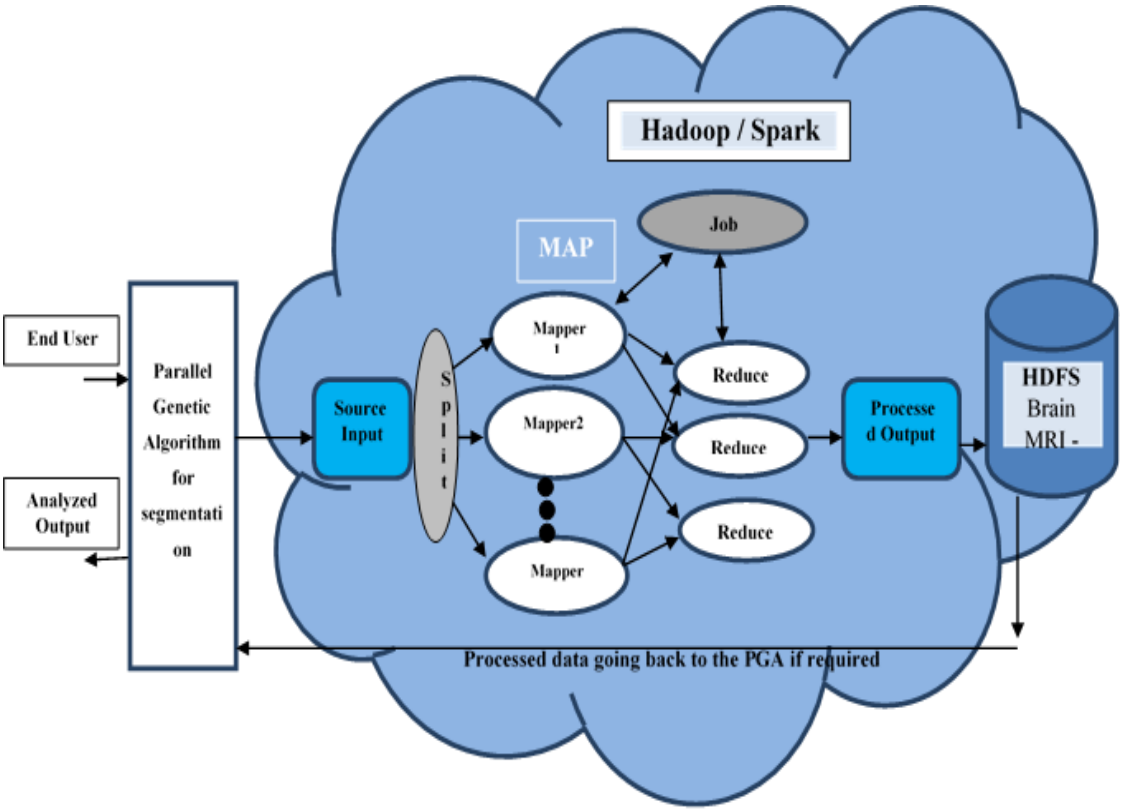

Figure 4.10 Proposed architecture for processing brain MRIs with multi cluster Apache Hadoop / Spark in cloud

Table 4.6 Performance Evaluation in Single Cluster Hadoop

Performance Evaluation in Single Cluster Hadoop using Map Reduce					
	Process Throughput in GB /Secs				
	2	**4**	**6**	**8**	**10**
1	12	15	17.2	20	23.4

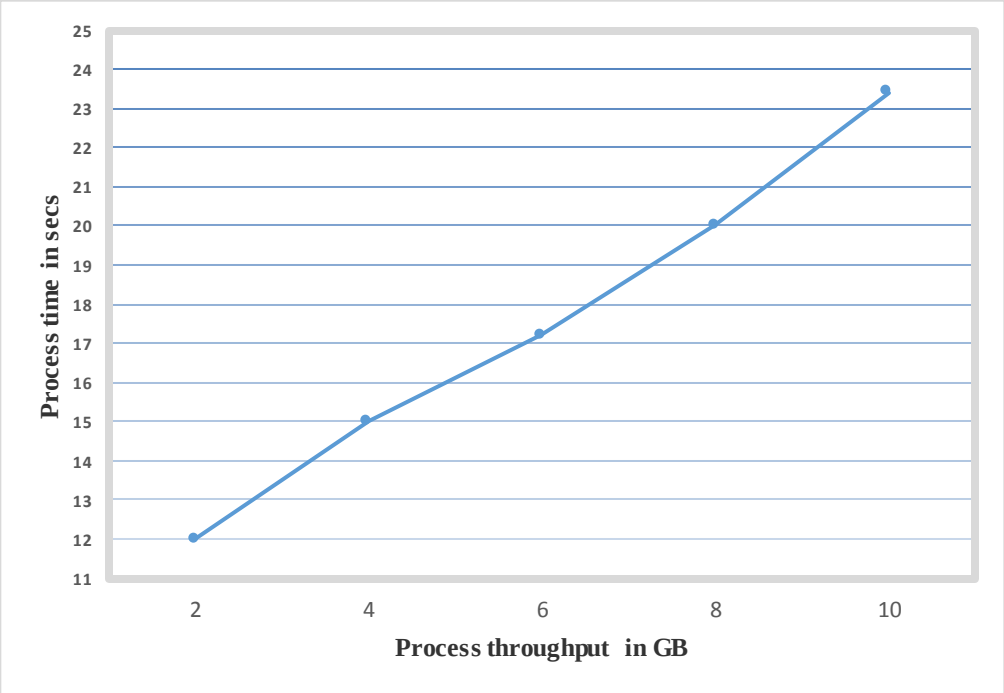

Figure 4.11 Process throughput in GB/sec in Single Cluster Hadoop

Table 4.7 Performance Evaluation in Multi Cluster Hadoop

Performance Evaluation in Multi Cluster Hadoop using Map Reduce					
Number of data nodes in a cluster	**Process Throughput in GB /Secs**				
0	**2**	**4**	**6**	**8**	**10**
2	13	16	16.4	17	18
3	13	15	15.1	15	15.5
4	14	15	15	14.8	14.6
5	14	14.6	14.4	14.1	14

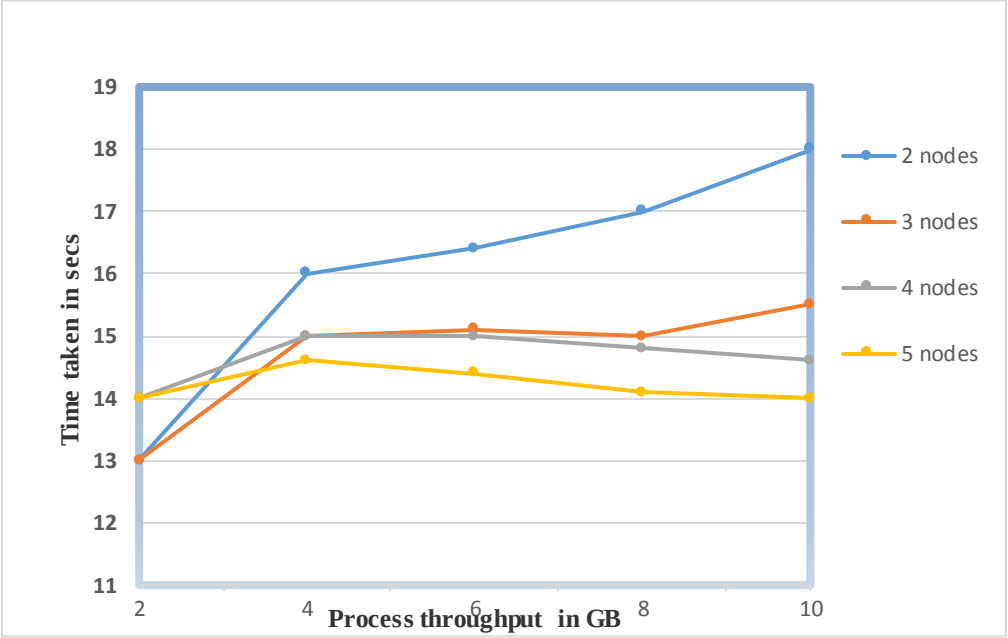

Figure 4.12 Process throughput in GB/sec in Multi Cluster Hadoop

When the same setup was used for Spark1.4.0 to execute the PGA, there is a drastic change in reduction of execution time to less than half of Hadoop taken for execution. Since Spark uses the memory for keeping intermediate reductions as discussed by Zhao, Ling, and Sun (2015) this reduction of execution time could be realized. Whereas Hadoop uses the disk for writing intermediate results that increase latency time for to and fro communications with a disk which origins more execution time.

Table 4.8 Performance Evaluation in Single Cluster Spark

Performance Evaluation in Single Cluster Spark					
	Process Throughput in GB /Secs				
	2	**4**	**6**	**8**	**10**
1	4	5	5	5	6

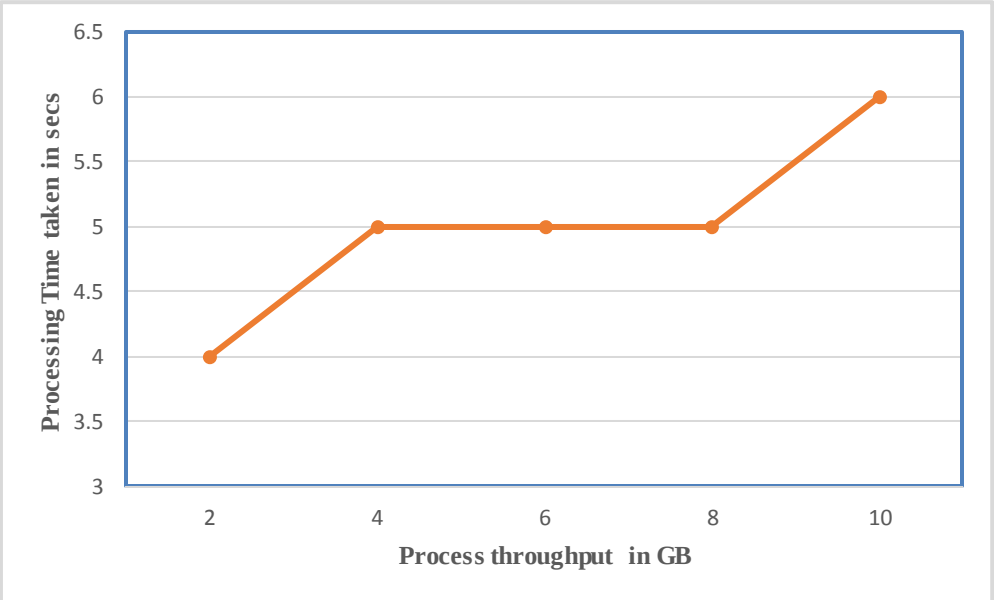

Figure 4.13 Performance Evaluation in Single Cluster Spark

Table 4.9 Performance Evaluation in Multi Clusters Spark

Performance Evaluation in Multi Clusters Spark					
Number of data nodes in a cluster	**Process Throughput in GB /Secs**				
	2	**4**	**6**	**8**	**10**
2	5	6	6	7	7
3	5	6	6	5	5
4	6	6	5	5	4
5	6	7	5	4	3

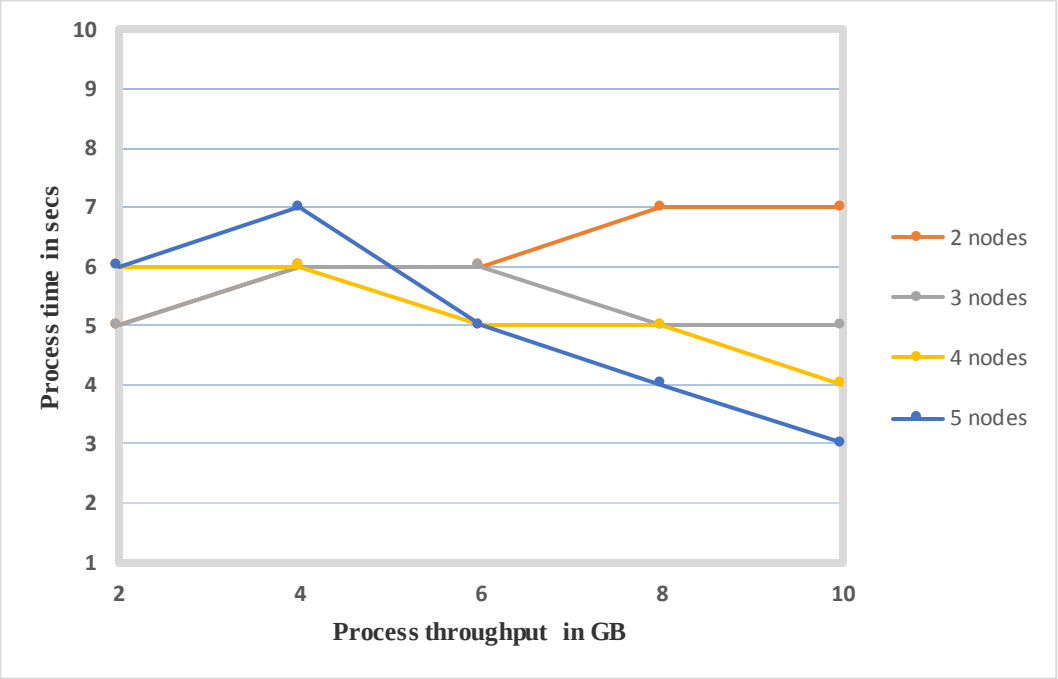

Figure 4.14 Process throughput in GB/sec in Spark1.4.0

4.8. ANALYSIS AND DISCUSSION

The objective of was to enable the interaction of several parallel subcomponents of an evolving population in the brain MR image in the cloud environment using Hadoop/Spark.

4.8.1. Increase in Brain MRI Segmentation Efficiency

Since PGA has shown extraordinary results in medical image processing, three different methodologies proposed by different authors have been tested with the given MRI brain images. Finally, the proposed methodology is tested with the same image data and compared with the results produced by the other three methods. We could see the improvement in the accuracy in segmenting the images in every algorithm. The algorithms are ordered based on the ascending order of values of accuracy they yield. In all these algorithms, the maximum generations were 1024 and the demes were small in size of eight.

In the first tested algorithm, Island Model where migration occurred among the adjacent demes alone, gives least value in accuracy among the four algorithms tested. Because of the early convergence of solution, the performance of PGA is not appreciated. In this model, individuals with optimum fitness value are transferred occasionally since there are not many options for competing individuals to take part. Moreover, the solution converged in 403 generations which are very earlier than the prescribed number of generations.

In the second tested algorithm, the individuals in each deme got exchanged with the individuals of other demes with the hybrid topology and with a

fixed migration rate. It is realized that the fitness of average population increases in the smaller demes defined. We could see some improvement in the accuracy because individuals' migrate freely there is an increase in chance for mutation to occur with the best individuals in other demes. But at the same time, due do the fixed migration rate which tends to be low, the demes absorbed the behaviour of independence and they could migrate to any fragment of the search space. But then again this migration attribute could not give any betterment in the results since the demes behaved as if they were in island model. It reveals that there is a probability of an increase in performance of PGA to a certain point of migration rate and after this point, the performance will decrease to the level of island model.

In the algorithm, where migration occurred at fixed intervals between processors we could find the restriction on the migration either before the stipulated time or after the stipulated time. It creates a major short fall in the performance of the algorithm since the controlled environment for the migration makes the algorithm work as a serial PGA. The individuals migration from different demes and different multiple processors in the parallel environment makes the algorithm as serial when there is a time delay in migration.

In the proposed methodology, we started with the initial population size of eight whereas the subpopulation size is s16. The number of neighbours with whom migration takes place is initialized as three to denote the immediate neighbours from where the expansion of the neighbours' network can take place based on the communication topology defined. Since the search space is broader because of the brain MR images, the number of generations

initialized as 1024. The fitness is found for each individual in a subpopulation using the objective function through which the migrant is picked for migration to the matching destination subpopulation through the crossover operation. The individual with lower fitness is replaced by the higher one.

Since the migration rate is calculated depends on the population size and the number of generations, it makes the algorithm dynamic. Since the division of population is related to the processing time at some point in time with respect to the communication topology and the dynamic increase in data size, the population size is taken as eight. The population size present in each node is estimated by dividing the total population size by the number of processors used.

The proposed algorithm eliminates island characteristic by providing the hybrid topology where each individual can travel to other best fitting demes among the processors. It will constitute the algorithm as a real time evolution. The algorithm avoids the migration to the extreme end subpopulation to mutate with best fitness individual to evade the consumption of time spending in travel. The two negative characteristics such as laziness of island model and unending process of searching for the best fitness individual of stepping stone are eluded with the average migration rate which makes the algorithm of down-to-earth life.

4.8.2. Increase in Performance in Cloud Setup

One of the main objectives of the proposed research is to provide a cloud environment where the processing of brain MRI can be carried out anywhere from the globe. To meet the objective by achieving the best

solution, we have run the proposed PGA in Hadoop as well as in Spark with single and multiples clusters in AWS. Instead of carrying out to evaluate the performance of proposed PGA for a single image, keeping the Big Data analytics in mind, the process was carried out for GBs of brain MRI data.

Performance evaluation was done for 2, 4, 6, 8 and 10 GB brain MRI data in a number of clusters starting from 1 to 5. Compared to Single cluster Hadoop, multi cluster Hadoop could perform faster with the size of data increases. In a single cluster Hadoop, 12 secs were taken to process 2GB of data whereas only 23.4 secs were taken to process 10 GB data. We are able to see that the processing time of PGA comes down even with the increase in data size in a single cluster setup. In Hadoop multi cluster setup, initially, there was either decrease or no change in the performance of proposed PGA with 2GB data. Since communication overhead between nodes increases due to the hybrid topology, there is a probability for either increase or no change in computation time.

But we are able to meet the objective with the increase of data size by experiencing the reduction in execution time in multi cluster Hadoop. As it is mentioned earlier, Hadoop works with MapReduce technique and meanwhile PGA also executes with parallel processing and reduction with migration topology, it is double benefited with respect to the huge amount of data in terms of execution time.

It is similar to the case of Apache Spark in which the intermediate storage takes place in the memory itself whereas it happens with a disc in Apache Hadoop. We could understand the power of Spark in terms of drastic reduction in execution time. In Spark, time taken for processing 10 GB of

data was 6 secs which are only 25% of the computing time taken by Hadoop to process the same volume of data. Even though the variation in processing time is less with respect to multiple clusters in Spark, as the data size increases, computing time also decreases radically. When we compare the processing time 14 secs and 3 secs in five clusters setup for 10 GB of data in Hadoop and Spark respectively, Spark consumes approximately 20% of the computing time taken by Hadoop.

So we are able to realize the power of parallelism with average migration technique integrated into Genetic Algorithm processed either in Hadoop or Spark which can be extended to support huge brain MRI data set termed as "Big Data" available globally in the cloud to be accessed from anywhere.