Parallelization of Genome Sequence Analysis Using Bigdata Approach

Sanjiva.S.G
Department of Information Science and Engineering
SDMCET, Dharwad, Karnataka, India

Leena.I.Sakri
Department of Information science and Engineering
SDMCET, Dharwad, Karnataka, India

Abstract- Genomic data refers to the ordering associate in nursing the DNA knowledge of an organism. They are utilized in bioinformatics for aggregation, storing and process the genomes of living things. Genomic knowledge usually needs an oversized quantity of storage and purpose-made computer code to investigate. The sequence could be an assortment of nucleotides or aminoalkanoic acid residues that area unit connected with one another. Speaking biologically, a typical DNA/RNA sequence encompasses nucleotides whereas a macromolecule sequence encompasses amino acids. A genome sequence is not associate in nursing finish in itself. A serious challenge still needs to be met in understanding what the ordination contains and the way the ordination functions. The previous is self-addressed by a mix of laptop analysis and experimentation, with the first aim of locating the genes and their management regions. This paper talks about the execution of the Hadoop, Map and decrease capacities to keep running in parallel and use the centers accessible in the virtual machines. This work embraces a generally acknowledged and exact Smith Waterman calculation algorithm for sequence arrangement and parallelization philosophy of map and reduce framework. Availability: The Saccharomyces cerevisiae S288 dataset and the respective links are available. Available @: https://downloads.yeastgenome.org/sequence/S288C_reference/

Keywords – Saccharomyces cerevisiae, parallelization, genomic data, genome sequence.

I. INTRODUCTION

Now a day the dimensions of information is being obtaining larger from quite whereas currently. From the dawn of your time to lower than a decade past (till 2003 precisely), world generated concerning 5 Exabyte of information. In 2012 in keeping with IBM, world knowledge grew up 2.7 Zettabyte, i.e., somewhere roughly five hundred times a lot of knowledge than all knowledge ever generated before 2003. Cisco survey that the, world knowledge on net can grow three times larger i.e., close to 4.8Zettabyte by the tip of 2015. One reason for knowledge growing larger is that, its unceasingly being generated by variety sources adore sensors, cc TV cameras, social media, etc. and by selection devices.

Much of that information equivalent to videos, photos, comments on social forums, reviews on various webpages is unstructured, which implies knowledge is not hold on ancient structured predened tables. Furthermore the information sources are incoming so quick that there is not even time to store it and apply logic to that. That is why the normal knowledge management and analytics tools alone do not alter IT to store, manage, method and analyze huge knowledge. i.e., big data.

The Big data specify to 'Data' whose size is on the far side the power of current technology to process, manage and capture the information among the advance time. We tend to live presently in information world. All over we tend to see solely information. So it's necessary to find out a way to store this information likewise as a way to method it. The answer to such a challenge is shifting progressively from providing hardware to provisioning a lot of manageable code solutions. Massive information in addition brings new opportunities and important challenges to industry and domain.

1. What is Hadoop?

In easy language, we are able to outline Hadoop as a framework for process and storage of big quantity of knowledge with cluster of product hardware. Another applicable definition would be as follows "Hadoop could be an open supply computer code given by Apache computer code foundation that allows distributed storage and process of huge datasets across clusters of servers." It has designed to scale up/down with great degree of fault

tolerance Hadoop was derived from Google's Map cut back and Google's classification system. Yahoo was creator in addition as major contributor and user of Hadoop across its business. Alternative major user embrace Twitter, Facebook, yank Airlines, Linked Inn, IBM, The big apple Times and lots of additional. The basic facet of flexibility of hadoop cluster is predicated on software's ability to discover and handle loss at the application layer. Apache Hadoop are often split into 2 core ideas particularly, HDFS (Hadoop Distributed File System) and MapReduce. To place it in straightforward terms, HDFS could be a technique for storing large information sets. MapReduce is technique for process information hold on HDFS. HDFS assumes nodes can fail. It achieves reliableness by replicating information across multiple nodes. Hadoop is build up by eco system of Apache comes resembling Hive, Zookeeper, Pig, that improves usability and extend the worth of Hadoop. The Hadoop changes its dynamics and social science of Big Data computing.

2. Understanding Map Reduce

MapReduce may be a programming model for process massive knowledge sets with a parallel, distributed rule on a cluster. Map scale back once let alone HDFS is accustomed handle massive knowledge. The basics of this HDFS MapReduce system that is usually brought up by Hadoop. The basic unit of knowledge, utilized in Map Reduce may be a (Key, value) combine. All kinds of structured and unstructured knowledge ought to be translated to the current basic unit, before feeding the info to Map Reduce model. Because the name suggests, MapReduce model comprises 2 separate routines, particularly Map-function and Reduce-function. This text can assist you perceive the step by step practicality of MapReduce model. The computation on associate degree input (i.e. on a group of pairs) in MapReduce model happens in 3 stages:

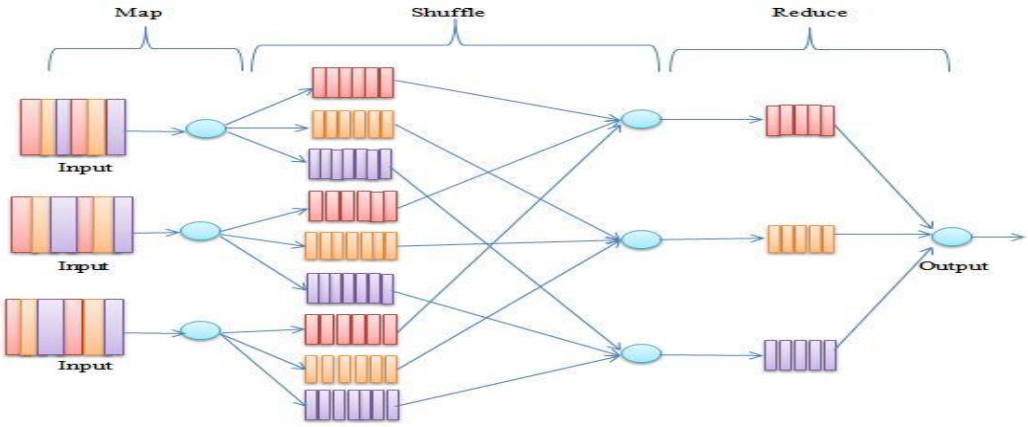

Fig1: Map Reduce Programming Paradigm

In the map stage, the plotter takes one (key, value) combine as input and produces any variety of (key, value) pairs as output. It's necessary to consider the map operation as homeless, that is, its logic operates on one combine at a time (even if in apply many input pairs are delivered to constant mapper). To summarize, for the map section, the user merely styles a map perform that maps associate degree input (key, value) combine to any variety (even none) of output pairs. Most of the time, the map section is solely wont to specify the specified location of the input worth by ever-changing its key. The shuffle stage is mechanically handled by the MapReduce framework, i.e. the engineer has nothing to try and do for this stage. The underlying system implementing MapReduce routes all of the values that area unit related to a personal key to an equivalent reducer. In the scale back stage, the reducer takes all of the values related to one key k and outputs any range of (key, value) pairs. This highlights one among the sequent aspects of MapReduce computation: all of the maps got to end before the scale back stage will begin. Since the reducer has access to any or all the values with an equivalent key, it will perform sequent computations on these values. Within the scale back step, the similarity is exploited by perceptive that reducers in operation on totally different keys are dead at the same time. To summarize, for the scale back section, the user styles a operate that takes in input a listing of values related to one key and outputs any range of pairs. Typically the output keys of a reducer equal the input key (in reality, within the original MapReduce paper the output key should adequate to the input key, however Hadoop relaxed this constraint). Overall, a program within the Map Reduce paradigm will contain several rounds (usually referred to as jobs) of various map and scale back functions, performed consecutive one once another.

3. About Sequencing:

Sequencing is that the method to see the ester or aminoalkanoic acid sequence of a polymer fragment or a super molecule. There are completely different experimental strategies for sequencing, and also the obtained sequence is submitted to completely different databases like NCBI, Genbank etc. There area unit completely different strategies and machines that may sequence genomes. In the 100,000 Genomes Project, deoxyribonucleic acid is sequenced by our partners at Illumina. One human order is often sequenced in a couple of day, although the analysis takes for much longer. DNA sequencing machines cannot sequence the total order in one go. Instead, they sequence the deoxyribonucleic acid briefly items, around one hundred fifty letters long. Every of those short sequences are termed a 'read'.

4. Sequence alignment:

Sequence Alignment or sequence comparison lies deep down of the bioinformatics that describes the method of arrangement of DNA/RNA or macromolecule sequences, so as to spot the regions of similarity among them. Its accustomed infers structural, practical and biological process relationship between the sequences. Alignment finds sameness level between question sequence and completely different information sequences. The rule works by dynamic programming approach that divides the matter into smaller freelance sub issues. It finds the alignment a lot of quantitatively by distribution scores. When a brand new sequence is found, the structure and performance are often simply expected by doing sequence alignment. Since it's believed that, a sequence sharing common ascendant would exhibit similar structure or perform. Bigger the sequence similarity, bigger is that the probability that they share similar structure or perform.

4.1 Methods of Sequence alignment:

a. Global Alignment: It was actually proposed by Needleman-Wunsch. It was actually suitable for closely related sequences of same length. Here the alignment is carried out from the beginning of the sequence till the end of the sequence in order to find out the best possible alignment between the two sequences.
b. Local Alignment: It was developed by Smith-Waterman. The sequences which are suspected to be similar or even dissimilar sequences can also be compared. It actually takes only the concerned regions or the functionality region for the comparison. It requires to main modifications, Firstly high negative scores for mismatches and then a value in the score matrix that is the tabulation matrix it becomes negative the values are reset to zero.

5. About Smith Waterman Algorithm:

The Smith-Waterman algorithmic program is a dynamic programming language were a info search algorithmic program developed by T.F. Smith and M.S. Waterman, And supported an earlier model suitably named Needleman-Wunsch once its original creators. The S-W algorithmic program implements a method known as dynamic programming that takes alignments of any length, at any location, in any sequence, and determines whether or not a best alignment is found. Supported these calculations, scores or weights square measure allotted to every character-to-character comparison: positive for precise matches/substitutions, negative for insertions/deletions. In weight matrices, scores square measure other along and therefore the highest evaluation alignment is according.
It is higher to the BLAST and FASTA algorithms as a result of it searches a bigger field of prospects, creating it an additional sensitive technique; but, a single pair-wise comparison between letters slows the method down considerably.
Instead of viewing a complete sequence directly, the S-W formula compares multi-lengthen segments, searching for whichever section maximizes the evaluation live. The formula itself is algorithmic in nature:

$$M\ i\ j = \text{Max}\{M\ i\text{-}1,\ j\text{-}1 + S\ i,\ j,\ M\ i,\ j\text{-}1 + W,\ M\ i\text{-}1,\ j\text{+}\ W, 0\}$$

The Steps that are involved in alignment are:

- Initialization of a matrix.
- Filling the matrix with appropriate Scores.
- Trace back analysis.

II. PROPOSED SYSTEM

2.1 Parallelized Hadoop Map Reduce

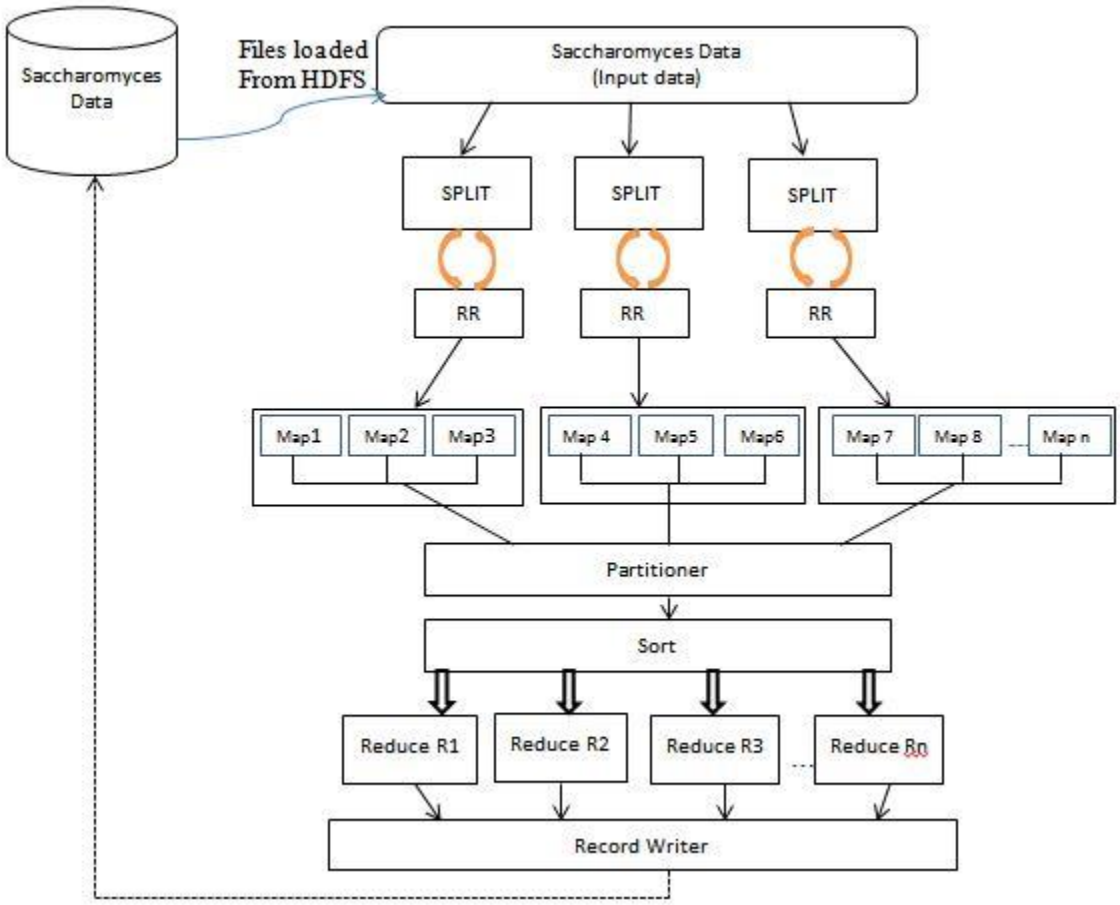

Fig 2: Parallelized Hadoop Map Reduce

Map reduce provides the framework for highly processing of data across clusters of commodity hardware. From the above fig the saccharomyces Input data is split into smaller chunks that are produced in parallel on cluster of nodes by programs called mappers. By default the HDFS block size is chosen as the split size and one map task is initiated for each HDFS block of input. Map tasks do not share anything between them and are run in parallel. Each block contains multiple map tasks may run on the same node in parallel. If the node has more than one map slot available and also has multiple blocks of the input file. Output of map tasks are portioned and sorted before presenting to reducers. Map reduce frame work takes care of the shuffle/short. The frame work guarantees that the keys are ordered and all the values for a particular key presented to the same reducer.

A reducer task takes the shuffle output and provides the results R1, R2, R3….. Rn by aggregating data from all the mappers. As many as output files are produced as the number of reducers are produced. Reducers also do not share anything with ordered reducers and hence run in parallel. The final output of the reducer is written into the record writer. Finally the record writer writes in the HDFS (Hadoop distributed file system).

III. EXPERIMENT AND RESULT

The framework condition utilized is windows 7 undertakings 64-bit, 8 GB ram, I-5 quad core processor working framework. Dot net framework 4.0 and C# 6.0 programming dialect is utilized for the proposed work and java programing dialect is utilized for the current Hadoop and led exploratory examination on following parameter for direct speed ups and undertaking finishing time and contrasted the proposed genome sequencing model and existing sequencing model.

To assess the execution of Smith Waterman parallel MapReduce SW-PMR correlation with the SW-Hadoop is considered. The organization of SW-Hadoop and SW-PMR is considered with one VM figuring hub. The baker yeast genomic database (Saccharomyces cerevisiae S288c) is considered for assessment. Tests are led utilizing a steady reference genomic sequences and four query sequences of fluctuated lengths are considered. The investigations directed with the reference and question genomic sequences are abridged. Considering the smith waterman arrangement calculation on the SW-PMR and SW-Hadoop bunches the aggregate time taken to execute the arrangement is observed. The time taken of the Map organize (which is appeared in Fig 3&6) and the Reduce arrange (which is appeared in Fig 4&6) alongside the aggregate time taken to finish grouping arrangement (which is appeared in Fig 5&6) is watched. In Fig 8 the aggregate time taken to adjust sequence by various map worker to changed query sequence estimate is appeared and in Fig 9 the aggregate time taken by various diminish specialist with shifted question grouping size is appeared. In analyze 1 (succession length of 1K) the accelerate accomplished for SW-PMR is around 12.5. For grouping arrangements i.e. in analyze 4 (succession length of 500k which is appeared in Fig 3) the speedup was seen to be 43.5. As it is observed that the question length expands the calculation time to adjust grouping additionally increment and furthermore the execution of SW-Hadoop seriously debases with increment in inquiry estimate when contrasted and the execution of the proposed SW-PMR. Fig 7 demonstrates that a normal accelerate of 23.5 is accomplished considering SW-PMR when contrasted with the SW-Hadoop.

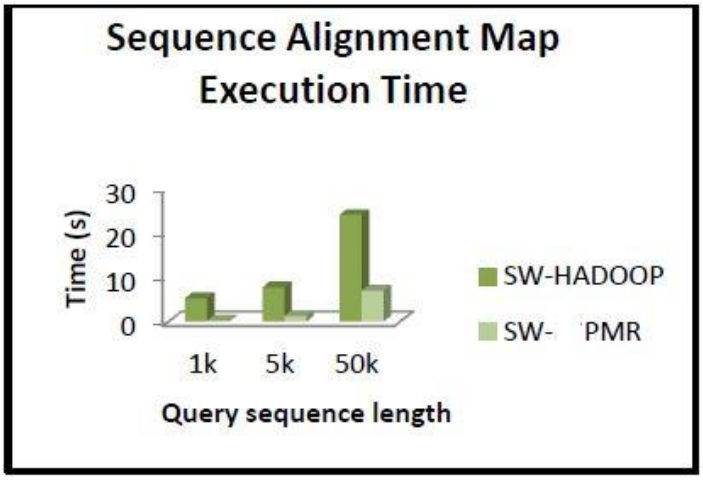

Fig3: Sequence alignment map execution

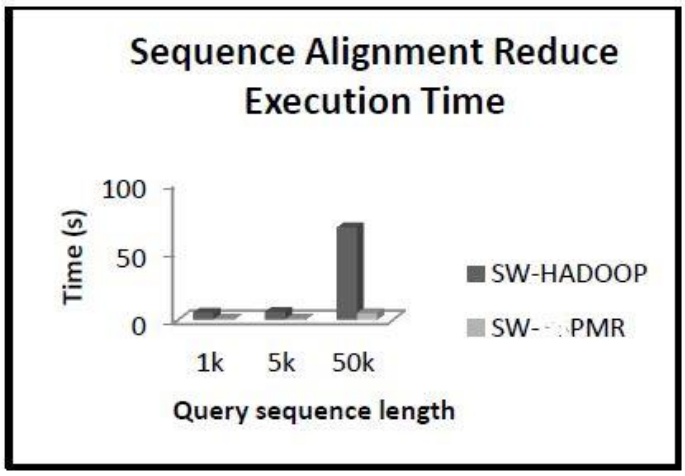

Fig 4: Sequence alignment reduce execution

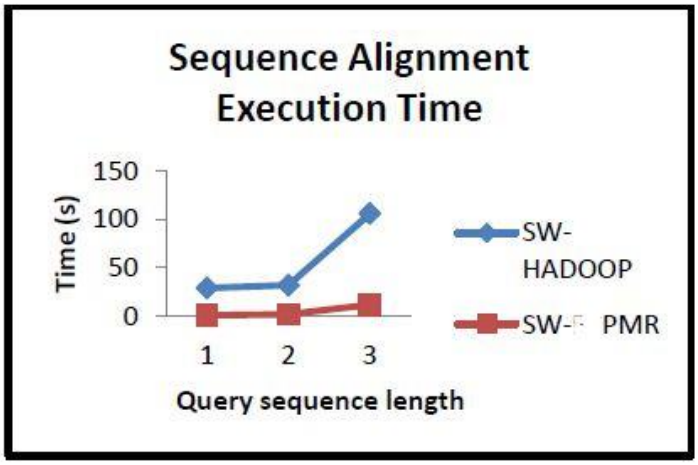

Fig 5: Sequence alignment execution time

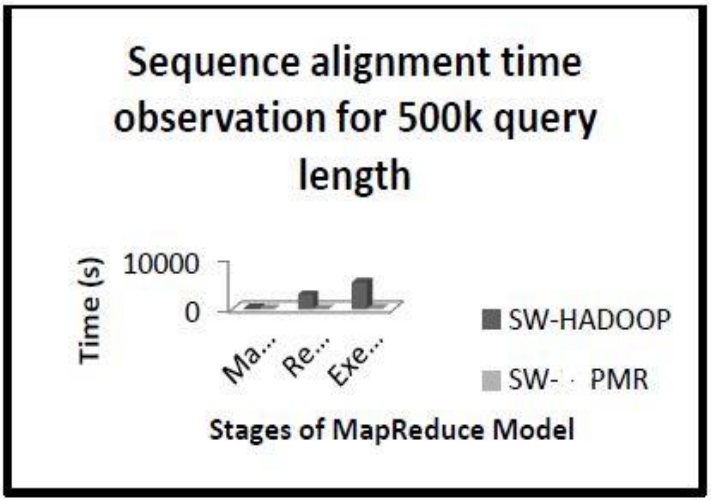

Fig 6: Observation time for 500k query length.

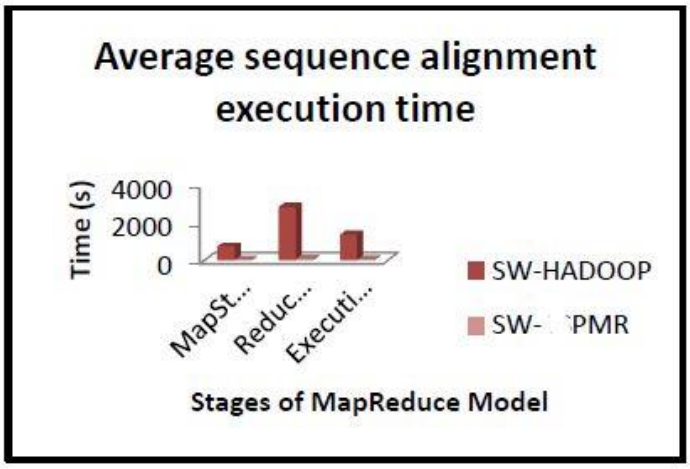

Fig 7: Average sequence alignment execution time

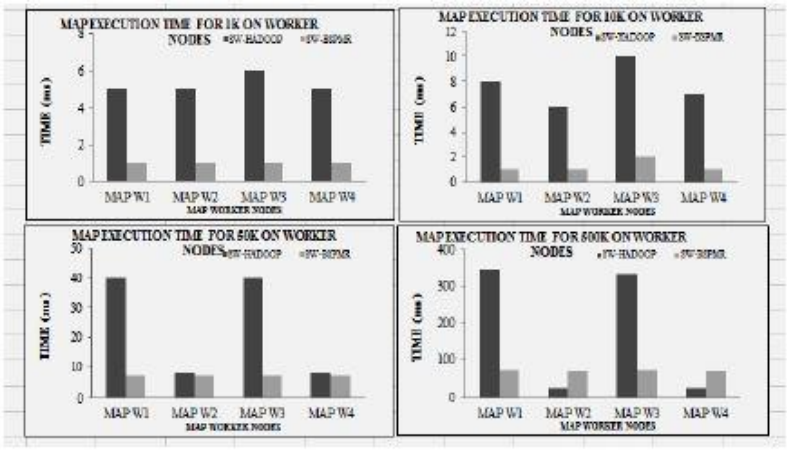

Fig 8: Map execution time for varied sequences.

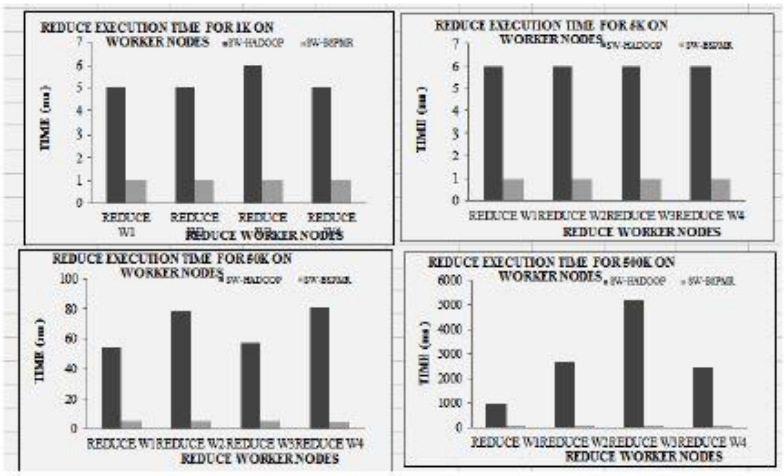

Fig 9: Reduce execution time for varied sequence

IV.CONCLUSION

This paper presents parallelized design execution utilizing hadoop framework for genome sequence alignment. MapReduce parallel programming speeds up the arrangement conveying the proposed architecture enhances functionality, scalability and adaptability of the framework. Furthermore, this can build the throughput of the arrangement because of undertaking parallelization. Expansive scale datasets gathered from the bio-informatics scientists is a challenging issue. To manage such enormous genome informational indexes Hadoop framework is most suitable. Time required for arrangement is likewise lessened, as HDFS makes information get to quicker. Whereas parallelized hadoop delineate is more proficient than the current architecture.

REFERENCES

[1] Kiran Menon, Kamalapriya Anala, Gokhale Trupti S.D, Neeru Sood, “Cloud Computing: Applications in biological research and prospects”, IEEE 2012.

[2] Biji C.L., Achuthsankar S. Nair, “Benchmark Dataset for Whole Genome Sequence compression”, IEEE 2017.

[3] Berard, A. Chateau, N. Pompidor, P. Guertin, A. Bergeron and K. M. Swenson, "Aligning the unalignable bacteriophage whole genome alignments," *BMC bioinformatics*, vol. 17, no. 1, 2016.

[4] Sumarish C. Purbarani, Hadaiq R. Sanabila, Anom Bowolaksono, Budi Wiweko, “A Survey of Whole Genome Alignment Tools and Frameworks based on Hadoop’s MapReduce”, IEEE 2016.

[5] Dr. Siddu P. Algur, Leena I. Sakri, “Parallelized Genomic Sequencing Model: A Big data approach for Bioinformatics application”, IEEE 2015.

[6] Miss. Anju Ramesh Ekre, Prof. Ravi. V. Mante, “Hadoop Based Clustering System For Genome Sequencing”, IEEE 2016.

[7] Y. Chen, W. Ye, Y. Zhang and Y. Xu, "High speed BLASTN: an accelerated MegaBLAST," *Nucleic acids research*, vol. 43, no. 16, pp. 7762-7768, 2015.

[8] S. Warris, F. Yalcin, K. J. L. Jackson and J. P. Nap, "Flexible, fast and accurate sequence alignment profiling on GPGPU with PaSWAS," *PloS one*, vol. 10, no. 4, p. e0122524, 2015.

[9] Q. Zou, X. B. Li, W. R. Jiang, Z. Y. Lin, G. L. Li and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Briefings in bioinformatics*, vol. 15, no. 4, p. bbs088, 2013.

[10] Divya D. Patel, Kavita R. Singh, "Genome Sequencing using MapReduce and Hadoop – A Technical Review", IEEE 2017.

[11] Syed Abdul Mutalib Al Junid, Mohd Faizul Md Idros, Abdul Hadi Abdul Razak, Fairul Nazmie Osman, "Parallel Processing Cell Score Design of Linear Gap Penalty Smith-Waterman Algorithm", IEEE 2017.

[12] O. Torreno and O. Trelles, "Breaking the computational barriers of pairwise genome comparison," *BMC bioinformatics*, vol. 16, no. 1, p. 1, 2015.

[13] Chinmayee Mohapatra, Leena Das, Siddharth Swarup Rautray, Manjusha Pandey, "Map-Reduce based Modeling and Dynamics of Infectious Disease", IEEE 2017.

[14] Vijay Naidu, Ajit Narayanan, "Needleman-Wunsch and Smith-Waterman Algorithms for Identifying Viral Polymorphic Malware Variants", IEEE2016.

[15] ChaoWang, Xi Li, Peng Chen, Aili Wang, Xuehai Zhou, Hong Yu, "Heterogeneous cloud framework for big data genome sequencing," IEEE/ACM transactions on computational biology and bioinformatics, vol. 12, no. 1, January/February 2015, pp. 166-178.

[16] Sullivan."Hadoop 2 vs. Hadoop 1 - HDFS and YARN," USENET http://www.tomsitpro.com/article/hadoop-2-vs-1,2-718.html, Sep. 23, 2015 [Dec. 13, 2016].