ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Analysis of DNA Data Using Hadoop Distributed File System.

Senthilkumar M[1]*, and Ilango P[2].

[1] School of Information and Technology and Engineering VIT University, Vellore, Tamil Nadu, India.
[2] Professor, School Of Computer Science And Engineering, VIT University, Vellore, Tamil Nadu, India.

ABSTRACT

The objective of this paper is to present how data can be parallel processed using hadoop distributed file system. Here, we are implementing the DNA data i.e., its sequence to find out the number of copies of the sequence, molecular weight and the number of copies of the molecular weight in the given data. Hadoop distributed file system is used to find out the repetitions of the given data so that the amount of original data can be reduced in size. Thus saving the space for storing a huge amount of data. Moreover, the replicas of a same data can be replaced with a single data and their replicas can be tracked with a unique identification and thus providing with efficient use of storages for a large set of data.

Keywords: DNA, Hadoop, file.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

There are many algorithms to process data and we have implemented using certain built-in examples of hadoop software framework. The data we have taken into analysis is DNA data i.e., its sequence. The sequence of DNA can be represented using four nucleobases namely (guanine, adenine, thymine, and cytosine) generally known as G,A,T,C respectively. We have use WordCount program to calculate the number of copies that is found in the sequence so that we can have a rough idea of how much amount of data is being repeated or replicated. Thus finding this will be the first step in processing the data[1]. Then the molecular weight of the processed dna sequence will be found. This will be again checked for the replication. Finally the number of duplicated molecular weight will be replaced by a single molecular weight thus huge amount of data can be reduced to certain extent[2]. This implementation is done using hadoop software framework and the correctness of the programs are concurrently checked in a java environment. Thus the main objective of using hadoop distributed files system is to reduce the size of the data by eliminating the repetitions or duplicates and providing a common data for the replicas[3].

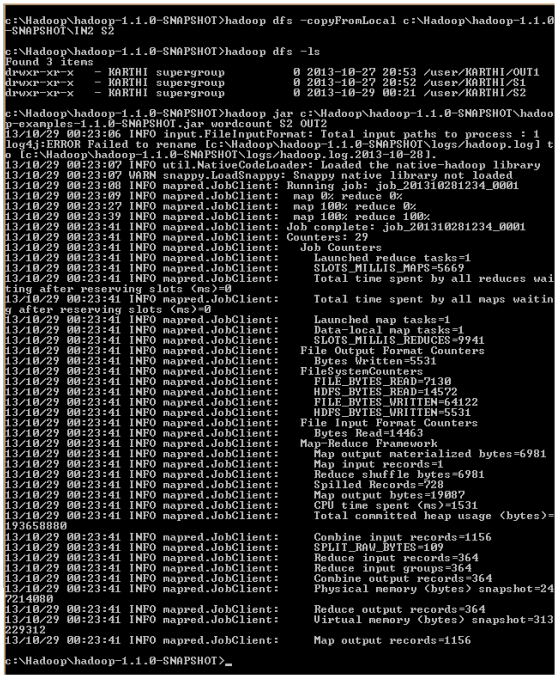
HDFS : Hadoop Distributed File System is sub-project of Apache Hadoop project. Hadoop is ideal for storing large amounts of data, in terms of TB (terabytes) and PB (petabytes). Hadoop uses hdfs as the storage system.

NameNode :It is the master-piece of Hadoop Distributed File System. It controls the directory of all files in the file system and tracks where the data is kept[4]. It does not store the data of these file itself.

DataNode : A datanode stores data in the Hadoop file system. More replication of data is found in a functional file system having more DataNodes. It serves read and write requests from the clients. With the instruction from the NameNode, it performs certain operations like block replication to replicate blocks, block creation to create blocks and block deletion to delete the blocks.

Block : A file internally split into one or more blocks which are stored in DataNodes. The default block size is 64 MB.

DESIGN AND IMPLEMENTATION:

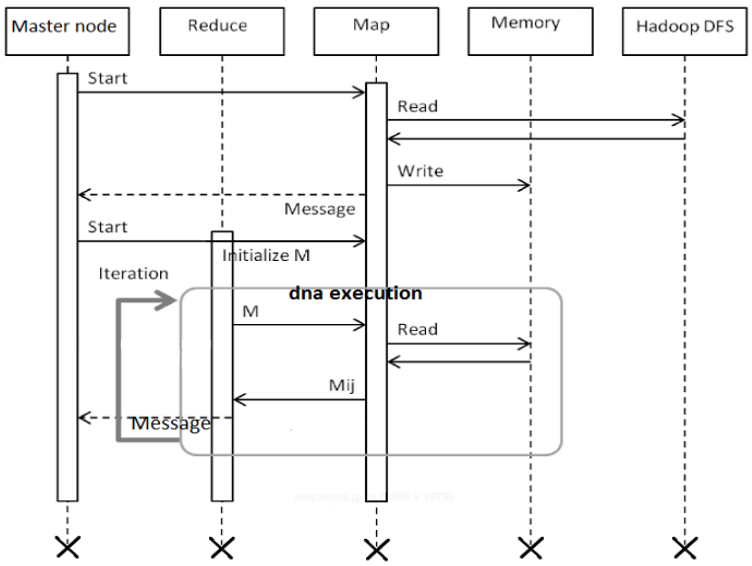


Implementation of the design concepts are done using hadoop software framework where a single node is created to perform mapreduce for the given data sets. Hadoop gives us the flexibility in choosing the various types of data so that any given data can be reduced in size if and only if the data size is huge in terms

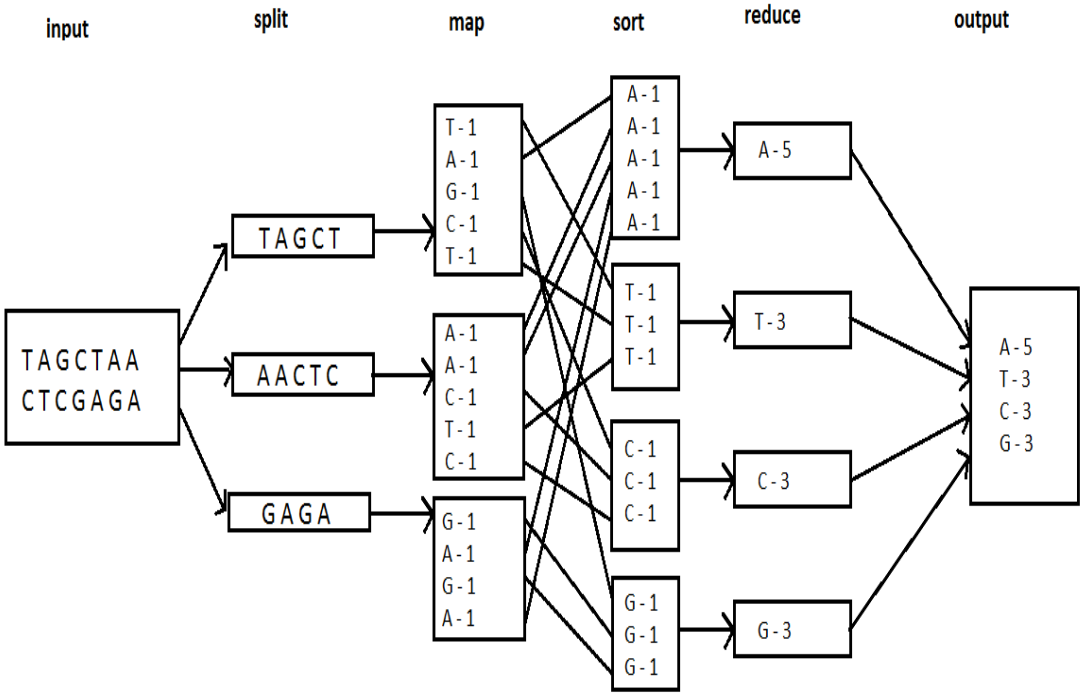
ISSN: 0975-8585

of petabytes or terabytes. Moreover the data with more replications find it more easier when implemented[5].

Sequence diagram : This gives the detailed view of the paper in a sequential manner. Here, the working of the HDFS is explained in the order of execution starting from the input data that we give to process till the final output file which is being created. The clear idea of how the whole program is executed is shown below.

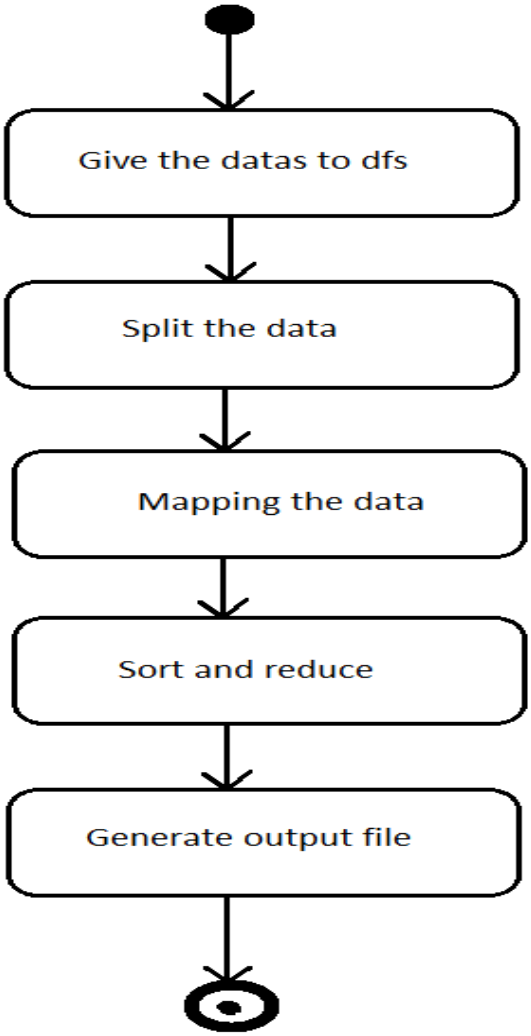


Data processing : The diagram shows the execution process of the input (DNA Data) for which "split" , "map" , "sort" , "reduce" is done by the hadoop distributed file system. This is implemented in hadoop software framework using two functions namely map and reduce. The map function takes (key, value) pair to perform filtering which is done in order to achieve a standard set of inputs for the processing of data in an efficient manner and sorting of the input file or data is aslo done so that it gives a clear look of how the data is arranged.The reduce function does the collection of the sorted data i.e. it groups the similar data so that the size is reduced and the output file is generated from the reduce which is the final result.



ISSN: 0975-8585

Activity diagram : This gives the flow of activities performed in order to get the final output in the required format. Here the steps to execute are explained. First, the data is given as input to the distributed file system. Then the data is split according to the program and mapping is done by the map function finally sort and reduce is done by the reduce function which generates the output file.



ALGORITHM

The algorithms used in the processing of the data are as follows:

i) Call dna function to perform the required operations and Read the input file until null.
ii)Compare a sequence with all the other sequences to find the repetition and Print the sequence and the number of occurrences. Also print the repeated sequence.
iii)Give the input file i.e., the dna sequence and Find the number of nucleobases in the file.
iv)Calculate molecular weight using formula m.w = (a* 313.2) + (t* 304.2) + (c* 289.2) + (g* 329.2) + 79.0 and Store the result of all sequence in an output file.
v)Read the output of the molecular weight calculation and Check the number of repetitions of the same molecular weights.
vi)Finally store the weight and the copies in an output file. Thus it gives the repetitions.

Working principles of Algorithm

DNA sequence check : This is to check the occurrence of the sequence in the input file.

ISSN: 0975-8585

i) Call dna function to perform the required operations.
ii)Read the input file until null.
iii)Compare a sequence with all the other sequences to find the repetition.
iv)Print the sequence and the number of occurrences. Also print the repeated sequence.

```
> run DNA_sequence_check
The sequence          CGAGTGGAATGGAATGGA   repeats
The sequence          GAATGGAATGGAATGGAA   repeats
>
```

Molecular weight calculation: This is to find the molecular weight of the sequence.

i)Give the input file i.e., the dna sequence.
ii)Find the number of nucleobases in the file.
iii)Calculate molecular weight using formula m.w = (a* 313.2) + (t* 304.2) + (c* 289.2) + (g* 329.2) + 79.0
4:Store the result of all sequence in an output file.

```
    {
        arr[i] = String.valueOf(line.charAt(i));
if(arr[i].equals("A"))
{
a++;
}
if(arr[i].equals("C"))
{
c++;
}
if(arr[i].equals("T"))
{
t++;
}
if(arr[i].equals("G"))
{
g++;
}
}
d= (a* 313.2) + (t* 304.2) + (c* 289.2) + (g* 329.2) + 79.0;

System.out.println("Molecular weight of"+line+"is:\t"+d);
b1.write(d+"\t");
i++;

        }
```

Copy number calculation: This is to check the number of replications of the molecular weight.

i)Read the output of the molecular weight calculation.
ii)Check the number of repetitions of the same molecular weights.
iii)Finally store the weight and the copies in an output file. Thus it gives the repetitions.

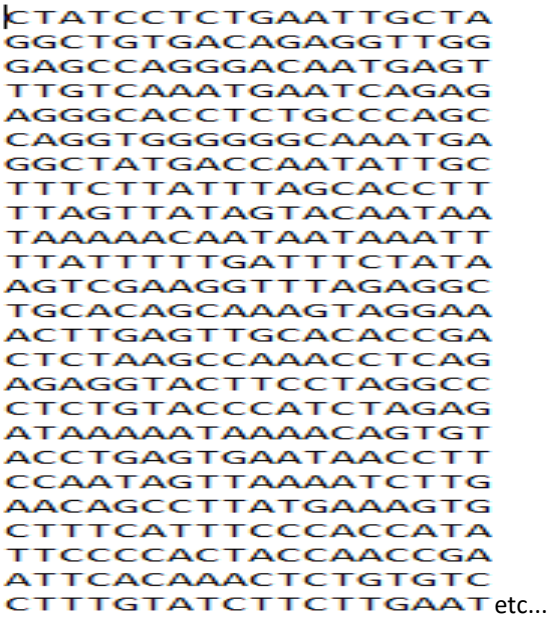
```
> run Molecularweight_calculation
Molecular weight of CTATCCTCTGAATTGCTA is: 5565.599999999999
Molecular weight of GGCTGTGACAGAGGTTGG is: 5776.599999999999
Molecular weight of GAGCCAGGGACAATGAGT is: 5738.6
Molecular weight of TTGTCAAATGAATCAGAG is: 5687.6
Molecular weight of AGGGCACCTCTGCCCAGC is: 5586.6
Molecular weight of CAGGTGGGGGGCAAATGA is: 5794.6
Molecular weight of GGCTATGACCAATATTGC is: 5639.6
Molecular weight of TTTCTTATTTAGCACCTT is: 5546.599999999999
Molecular weight of TTAGTTATAGTACAATAA is: 5661.599999999999
Molecular weight of TAAAAACAATAATAAATT is: 5647.599999999999
Molecular weight of TTATTTTTGATTTCTATA is: 5600.599999999999
Molecular weight of AGTCGAAGGTTTAGAGGC is: 5744.6
Molecular weight of TGCACAGCAAAGTAGGAA is: 5706.6
Molecular weight of ACTTGAGTTGCACACCGA is: 5624.6
Molecular weight of CTCTAAGCCAAACCTCAG is: 5553.599999999999
Molecular weight of AGAGGTACTTCCTAGGCC is: 5640.6
Molecular weight of CTCTGTACCCATCTAGAG is: 5575.6
Molecular weight of ATAAAAATAAAACAGTGT is: 5688.599999999999
Molecular weight of ACCTGAGTGAATAACCTT is: 5623.6
Molecular weight of CCAATAGTTAAAATCTTG is: 5622.599999999999
Molecular weight of AACAGCCTTATGAAAGTG is: 5672.599999999999
Molecular weight of CTTTCATTTCCCACCATA is: 5485.599999999999
Molecular weight of TTCCCCACTACCAACCGA is: 5489.599999999999
Molecular weight of ATTCACAAACTCTGTGTC is: 5574.599999999999
Molecular weight of CTTTGTATCTTCTTGAAT is: 5586.599999999999
Molecular weight of TATTGTCTTCCGAGGTTT is: 5627.599999999999
Molecular weight of TCATTCCAAATAGAATTT is: 5597.599999999999
Molecular weight of GCTGATTCATGTTAATTA is: 5644.6
Molecular weight of ATGTTAAGGGTCACATGG is: 5719.6
Molecular weight of TGAAGAAATGGGGTTGTT is: 5774.6
```

Program Explanation:

ISSN: 0975-8585

DNA sequence checking for repetition:

This phase involves in reading the input file which contains DNA sequence. The purpose of this program is to find the number of repeating DNA sequence (The Order in which the nucleotides are present in the DNA molecule).The nucleotides are nothing but certain letters in our sequence we use letters such as G, A, C and T. The DNA sequence is composed of specific letters while comparing one sequence of dna letter with forthcoming we could easily identify the repeated sequence. The process involves in checking the first sequences of nucleotides with next set of nucleotides until we have compared all the sets in the given data until we identify the repetition. Thus after the long process f comparing sequences has been finished we will have to eliminate the repeating values because we just need to have the count of similar data instead of having the exact replica. when all these work is done we will have a set of no-replicating DNA sequence. This non-repeating DNA sequence file is given as input for the next program.



Finding the molecular weight:

The file with no repetition of DNA sequence is give as input for this program. Where the molecular weight of each sequence is calculated and among them whose molecular weight appears to be same their DNA sequence is found. The molecular weight is calculated by adding the molecular mass of each nucleotide (G, A, C and T) present in the given sequence. when finding this molecular weight is used to find copy number of DNA molecule. Therefore, file with molecular weight for each DNA sequence is arrived as result in a file.

Copy number calculation:

Thus the final phase of program is to give total number of copied DNA sequence based on molecular weight that has given from previous program. To explain in detail the copy number is maximum probability that same type of sequence of DNA can be expected to arrive. The formula involved performs division of constant number by each DNA sequences molecular weight which is gained from the previous programs output file. Thus the final result of how many copies of same DNA sequence is there in a data of DNA.

ISSN: 0975-8585

```
5600.59999999999        8
5601.599999999999985    1
5601.59999999999        6
5602.6   5
5603.59999999999        3
5604.59999999999        3
5605.59999999999        1
5605.6   3
5606.59999999999        9
5607.59999999999        9
5607.6   2
5608.599999999999985    1
5608.6   4
5609.59999999999        6
5610.59999999999        6
5611.6   2
5612.59999999999        1
5612.6   2
5613.59999999999        3
5614.6   3
5615.59999999999        2
5615.6   8
5616.59999999999        7
5617.6   4
5618.59999999999        5
```

Future Enhancements:

Right now we are executing this mapreduce functionality using a single node to perform analysis on DNA data. This functionality can be further improved when executed under two or more nodes where the one node act as master and rest will act as slave. Thus it results in sharing processor, resource etc leading to greater performance and less time consuming than single node. Thus when implement across an organisation which leads in reducing and easy handling of large amount of data. It results in distributed storage across various organisations for computational work capabilities. It has a capacity to store large volumes of data resulting in high throughput. Since use of large block size it helps in best possible way of manipulating gigabytes,petabytes of data. The mapreduce framework is basically a distributed computing framework which allows parallel processing of work among data of numerous sizes

CONCLUSION

Thus by implementing this concept of mapreduce and processing of data using HDFS leads in saving a huge amount of data space. Based on the size of data used in processing, the amount of data space saved will also increase drastically. These concepts can also be implemented on various kinds of data. The mapreduce technique focus on satisfying only the business needs rather than worrying about how it would fit in with distributed system complications. In order attain maximum effort of faster execution of a job, Mapreduces splits or decomposes jobs as Map and reduce tasks and maintains schedules for remote execution across any platform . Hadoop initiates a new era of processing of data in modern technology.

REFERENCES

[1] MapReduce for Data Intensive Scientific Analyses ,Jaliya Ekanayake,shrideep Pallickara,Geoffrey Fox,ESCIENCE '08 Proceedings of the 2008 Fourth IEEE International Conference on eScience ,Pages 277-284 ISBN: 978-0-7695-3535-7

[2] Iris recognition on Hadoop: A biometrics system implementation on cloud computing Shelly ,Raghava, N.S. Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference onDate of Conference,15-17 Sept. 2011

[3] McKenna A et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data."Genome Research, vol. 20, pp.1297{1303, 2010.

[4] CloudDOE: A User-Friendly Tool for Deploying Hadoop Clouds and Analyzing High-Throughput Sequencing Data with MapReduce Wei-Chun Chung,Chien-Chih Chen,Jan-Ming Ho,Chung-Yen Lin,Wen-Lian Hsu,Yu-Chun Wang,D. T. Lee,Feipei Lai,Chih-Wei Huang,Yu-Jung Chang mail ,June 04, 2014

[5] 'Big data', Hadoop and cloud computing in genomics ,Aisling O'Driscoll,Jurate Daugelaite,Roy D. Sleator , Journal of Biomedical Informatics 46 (2013) 774–781

[6] Pradeep Mantha, Andre Luckow, and Shantenu Jha. Pilot MapReduce: An Extensible and Flexible MapReduce Implementation for Distributed Data, 2012. Accepted to MapReduce workshop 2012 (HPDC12).

[7] Michael Cardosa, Chenyu Wang, Anshuman Nangia, Ab-hishek Chandra, and Jon Weissman. Exploring MapReduce efficiency with highly-distributed data. InProceedings of the second international workshop on MapReduce and its appli- cations, MapReduce '11, pages 27–34, New York, NY, USA, 2011. ACM.