Abstract- The sequence of character is defined as text data. Text data become a big data and does not have any specific format and only can be processed using the original Hadoop. Hadoop is an Apache software project that includes challenging subprojects such as the MapReduce implementation and the Hadoop distributed file system (HDFS). Amazon web Service(AWS) provides virtual Cloud computing Service such as storing data using S3 service and processing big data using EMR service. Amazon Elastic MapReduce (EMR) uses the original Hadoop as a processing environment to its Cloud Computing services. Hadoop provides a improved version such as H2Hadoop, which is provide the data processing service for improving Hadoop performance. In this paper process to find the sequence of jobs in text data using H2Hadoop and Amazon EMR and also go through the difference between Amazon EMR and H2Hadoop, but that shows the H2Hadoop performance is more dynamic the Amazon EMR.

Keywords- Amazon EMR, H2Hadoop, Big Data, HDFS.

I. INTRODUCTION

Different Cloud computing service are provided basically by the Amazon Web Service and it is most common environment, such as S3 for storing data, EC2 for providing resizable compute capacity, EMR for processing BigData[1]. These services operate from 13 geographical region across the world. Text data is defined as sequence of character that have no specific pattern, it may be symbols, letters, numbers. It might be real and readable words usually have similar jobs that are frequently adopted to the data by its users. Also many of these data sets are unstructured, and required fast and simple processing. To increase the performance of the existing Hadoop and MapReduce algorithm, it is crucial to develop an algorithm based on the type of data sets and requirements of the jobs.

Sequence data analysis or pattern mining applications usually depends upon complex and huge data storing and processing adequacy such as DNA sequence aligning in the DNA data sets[3]. Current DNA sequencing machines can sequence millions of short(25 to 500bp) reads from random positions of the genome. Some of the sequence data sets are readable by Humans but by using traditional processing techniques it becomes tedious for understanding and processing.

Now a days Mobiles Phones are rapidly emerging as the ultimate multimodal sensor of human dynamics and behaviors. Equipped with GPS, bluetooth, accelerometers, cameras, and microphones, current phones have the potential of tracing multiple forms of the data at scales previously unattainable.

To explore structured, semi-structured or unstructured data new avenues have been opened for possibility of open source and commercial Cloud Computing parallel processing platforms.Before knowing more information, it is required to know about the Big Data and Hadoop.

II. INFORMATION ON BIG DATA

Big Data are the Datasets that expands so larger that they become awkward to work with using on-hand database management tools[2]. They are many points to explain the BigData with
traditional data such as data size, content collection and processing. The BigData having complex feature that means BigData cannot be processed using traditional processing techniques in a tolerable processing time. We can produce large amount of data by entering the age of big data through our everyday activities. BigData presents a real opportunity in identifying actionable insights from information. The major role behind the size of digital evidence is the massive increase in the size of availability of digital data[2]. The 5Volumes of BigData are: 1.Data Volume, 2.Data velocity, 3.Data variety, 4.Data veracity, 5.Data value.

III. HADOOP FRAME WORK

Hadoop is an open-source software framework that was developed by Doug cutting and written in Java for distributed storage and distributed processing. It is an apache framework and to processing and analysis of BigData it processing solutions the main components of Hadoop are 1)HDFS Hadoop distributed file system 2)MapReduce. HDFS is an open source implementation of the Google file system[5]. Although it appears as a ordinary file system, its storage is actually distributed among different data nodes in different clusters. HDFS is built on the principle of the master–slave architecture. The Master node also called as Name node it provides data service and access permission to the Slave nodes. Slave node is also called as Data node it serve as storage for the HDFS. Large files are allotted and further divided among multiple data nodes. The Map processing jobs located on all nodes are operated on their local copies of the data. It can be observed that the Name node Processing jobs located on all nodes are operated on their local copies of the data. It can be observed that the Name node stores only the metadata and the long information while the data transfer to and from the HDFS is done through the Hadoop API.

MapReduce:

To processing the large amount of metadata on group of computers with erroneous and weak communication links by using a parallel programming model that is MapReduce and it is based on purpose of using MapReduce technique is, to move computations to data nodes, rather than bringing data to computation node and as long as synchronized and shared global states are not required most of parallel application can be implemented in MapReduce.

MapReduce Allows the data processing and it can be done in two stage one is map stage and another one is Reduce stage. By using Map stage the Data is divided sets of keys and value pairs and their objects are processed in parallel and with parallel number that matches the node number dedicated as slaves.

This processing generates intermediate key and value pairs that are temporary and can later be directed to Reduce stage. The processing is carry on parallelly , within the Map stage or Reduce stage . The Map and Reduce stages appear in a sequential manner by which the Reduce stage starts when the Map stage completed. The data are divide set of key and value pairs and their objects are processed I parallelly by Map stage with a parallel number that matches the node number dedicated as Slaves. This process generates intermediate Key and value pairs that are temporary and later it is directed to reduce stages.

IV. AMAZON WEB SERVICES AND H2HAADOOP

Amazon EMR

Once of the most common web service provider is the Amazon Web Service AWS[2]. This service provides distributed storage services S3, it stores data in different node in cloud and also AWS provide a distributed processing service, it is Elastic Map Reduce (EMR) service, in cloud environment it process the MapReduce job, users who don’t have the ability to create their own Cloud cluster, S3 and EMR are work together to providing good solution for Hadoop and BigData analysis.

Distributing process is provided to the BigData by using Amazon EMR. Users have to create a cluster in the AWS and complete some configurations such as the number of nodes, hardware
capability of the node, and a couple of directories to retrieve and store data etc. The Amazon EMR provides different versions of Hadoop to be selected by the user during the creation of the group of nodes and it also uses the native Hadoop to process MapReduce jobs. Instead of creating your own cluster in terms of hardware. Using Amazon EMR makes the process of analyzing the data much easier. It cost is more because of the distribution of data between nodes that are located in different rack on the network.

**H2Hadoop**

H2Hadoop is a developed framework that works on top of Hadoop to process MapReduce jobs. To speed up the processing time to provides a solution it reducing the data input size from HDFS. By reading specific data sets that carry the target data it eliminates data redundancy[1].

A absolute analysis about H2Hadoop architecture and development is proposed in this paper to store the metadata of the executed job is the main idea behind H2Hadoop. Using Common Job Blocks (CJB) tables is stored in the name node and metadata of the similar jobs are stored in Name node and also Name node helps to direct future jobs to specific Data node that carry the required data sets. Data follows the concept of write-once and read-many, so there is no ability to do any changes in the source file in HDFS[1].

---

**V. IMPLEMENTING FINDING SEQUENCE JOB ON AMAZON EMR AND H2HADOOP**

When we have complex and unstructured data sets such as DNA it is not so easy process using traditional data analysis tools to finding sequence in text data sets. In this section we discuss the job implementation in Amazon EMR and H2Hadoop and compare between these frameworks. Some sequences that have been manipulated and founded, shown chromosomes in the DNA data sets are as follows. For example sequence1(GGGGCAGGGG) shows the protein that is located in all chromosomes. We have to search in all data blocks or chromosomes, each time when we want to find a sequence or a subsequence of it.

In contrast to that sequence 2 (AAGACGG TGG TAAGG) is located chromosome 1 and 8. We only have to search in these specific blocks or chromosomes each time we want to find a sequence or a subsequence of it.

**Table1:SOME SEQUENCE LOCATIONS DNA CHROMOSOMES**

<table>
<thead>
<tr>
<th>Sqn o</th>
<th>Common feature (Sequence)</th>
<th>Block Name/ID (Chromosome Name)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sq1</td>
<td>GGGGCGGGG</td>
<td>In all Chromosomes</td>
</tr>
</tbody>
</table>

---

Fig 1. Architecture of H2Hadoop
Under different settings in the group of nodes, we implement to finding sequence of jobs in the DNA data sets in Amazon EMR and H2Hadoop. We have the same amount of nodes in both environments, but nodes in Amazon have higher capabilities in terms of hardware than H2Hadoop cluster. Compare to Amazon EMR, H2Hadoop gave a higher performance.

**Task Assignment of the DataNode**

A healthy balance between disks processing on various DataNodes and Name node is restore the failed operation on remote blocks. Data Locality is achieved by Cloud file distribution, in each machine the file processing is done locally and any failed operation reads from the blocks and recovered through block replication. After launching a Job, Mappers and Reducers are immediately selected by Job tracker.

A client operation on HDFS has network file transparency and the distribution of blocks deferent machine across the cloud is transparent to the client. HDFS is oriented towards Hardware transparency. Processing DataNodes can be commissioned or decommissioned dynamically without affecting the client.

**VI. RESULTS AND EVALUATION**

The experiments show that Compare to Amazon EMR environment, H2Hadoop provides more positive results. H2Hadoop provides less data read in size and also reduces many related factors to the jobs, such as CPU processing time, number of read operations, and the data read size in bytes. The number of read operations is the most important topic in H2Hadoop, which means that number of blocks that are read by the system and also another important role is added that is block data size.
Figure 3 shows one of the results, which is the number of read operations in Amazon EMR compared with H2Hadoop. H2Hadoop reads data from HDFS and Amazon EMR read data from S3. In every sequence, Amazon EMR, having same number of read operation but there is a difference in read operation in h2hadoop based on how frequent the sequence exists in the DNA. By using H2Hadoop, the number of read operations for sequence 2 was reduced to 15 times, which increases the efficiency by 72.2%. The number of read operation is more in H2Hadoop compare to Amazon EMR, which is less read operation.

Fig 4. Same jobs for Hadoop and CPU processing time in Amazon EMR

There is huge difference between the CPU processing time for H2Hadoop and Amazon EMR for example %. In sequence 1, CPU processing time in H2Hadoop is less compare to the Amazon EMR, read the all data sets each time in sequence 2, CPU processing time is 669 seconds in Amazon EMR, but 50 seconds in H2Hadoop, which speeds up the processing performance by 92.5. The process of retrieving data costs more time then H2Hadoop because it is a private cloud and Amazon S3 stores data in nodes.

The third comparison, the number of bytes read by the System or data read size. the comparison between H2Hadoop and Amazon EMR for the sequences show as fig 4.
In sequence1, data read size is almost the same in both environments, in sequence2 Amazon EMR more than H2Hadoop by almost four times.

VII. CONCLUSION

This paper discusses the implementation of Sequence job in two different environments using Hadoop MapReduce technique which are Amazon EMR and H2Hadoop. The implemented jobs read real text data sets from HDFS in H2Hadoop and from S3 in Amazon EMR. Comparing with Amazon EMR and H2Hadoop, it also shows more efficient work result. So, it improves the related factors to data size, such as CPU processing time, number of read operations and the input data size of MapReduce job.

REFERENCES