Trends in Engineering and Technology (NCTET-2K17)
International Journal of Advanced Engineering Research and Science (IJAERS) *Special Issue-3*
https://dx.doi.org/10.22161/ijaers/nctet.2017.9 *ISSN: 2349-6495(P) | 2456-1908(O)*

Improving Hadoop Performance by using H2Hadoop MapReduce workflow in Text Datasets

C REKHA[1], A.REVATHI[2]

[1]PG Scholar, Department of CSE, S.V. College of Engineering, rekhachejarla.123@gmail.Com
[2]Assistant Professor, Department of CSE, S.V. College of Engineering, revathi.a@svcolleges.edu.in

ABSTRACT:-- Cloud Computing use Hadoop structure for managing BigData in parallel. Hadoop has certain constraints that could be mauled to execute the work profitably. These restrictions are by and large a prompt outcome of information locale in the package, occupations and errands organizing, and asset parcels in Hadoop. Competent asset parcel remains a test in Cloud Computing MapReduce stages. We propose H2Hadoop, which is an upgraded Hadoop arrange for that reduces the figuring cost related with BigData examination. The proposed working in like way keeps an eye on the issue of preferred standpoint assignment in adjacent Hadoop. H2Hadoop gives an overwhelming reaction for "content information, for example, discovering DNA strategy and the subject of a DNA gathering. In like way, H2Hadoop gives a capable Data Mining approach for Cloud Computing conditions. H2Hadoop building impacts on NameNode's capacity to pick organizations to the TaskTrackers (DataNodes) inside the package. By adding control parts to the NameNode, H2Hadoop can commendably incite and dole out errands to the DataNodes that contain the required information without sending the work to the entire package. Separating and neighborhood Hadoop, H2Hadoop lessens CPU time, number of read operations, and another Hadoop parts.Index Terms— BigData, Cloud Computing, Hadoop, H2Hadoop, Hadoop Performance, MapReduce, Text Data

1. INTRODUCTION

Parallel get ready in Cloud Computing has created as an interdisciplinary research zone on account of the heterogeneous nature and broad size of data. Making an understanding of sequential data to imperative information requires huge computational power and viable counts to perceive the level of comparable qualities among different game plans. Back to back case mining or data examination applications, for instance, DNA game plan modifying and subject finding generally speaking require significant and complex measures of data planning and computational capacities. Capably concentrating on and arranging of computational resources is required to handle such complex issues. Though, a bit of the educational accumulations are discernable by individuals, it can be astoundingly mind boggling to be fathomed and arranged using standard taking care of strategies. Availability of open source and business Cloud Computing parallel planning stages have opened new streets to examine sorted out, semi-composed or unstructured data. Before we go any further, it is imperative to describe certain definitions that are related to BigData and Hadoop.

BigData Concepts

There are different techniques for portraying and separating BigData and the standard information, for example, information measure, substance, collecting and prepare. Huge information has been depicted as wide illuminating aggregations that can't be dealt with utilizing normal arranging strategies, for example, Relational Database Management Systems, in an unremarkable prepare time. BigData is either a social database (Structured, for example, securities exchange information or non-social database (Semistructured or Unstructured, for example, online frameworks organization information or DNA educational records.

The 4V's of BigData are 1) Volume of the data, which proposes the data survey. Some of affiliations' data stockpiling is about Zetabyte. 2) Velocity, which proposes the speed at which the data is made. 3) Varity of the data, which proposes the data graphs that unmistakable applications coordinate, for instance, development data, numeric data or twofold data. 4) Veracity of the data, which derives the shortcoming of the status of the data or how clear the data is to these applications.

Organized challenges in BigData have been evaluated in past

https://dx.doi.org/10.22161/ijaers/nctet.2017.9 ISSN: 2349-6495(P) | 2456-1908(O)

research [9] and they are depicted as particular burdens, for instance, the physical stockpiling, that stores the Big Data and lessening the overabundance. In like route, there are different challenges, for instance, the path toward isolating the information, cleaning data, data mix, data aggregation, and data outline. Since BigData has these issues, it needs such a circumstance or framework to work through these challenges. Hadoop, which works with BigData sets, is a structure that most affiliations use to oversee BigData remembering the ultimate objective to whipping data challenges. Hadoop is an Apache open-source programming structure that is made in Java for scattered stockpiling and passed on managing. It gives answers for BigData managing and examination. It has a report structure that gives an interface between the customers' applications and the close to record structure, which is the Hadoop Distributed File System HDFS.

Hadoop passed on File System ensures strong sharing of the points of interest for productive data examination. The two fundamental segments of Hadoop are (i) Hadoop Distributed File System (HDFS) that gives the data persevering quality (passed on stockpiling) and (ii) MapReduce that gives the structure examination (appropriated organizing). Subordinate upon the pick that "moving number towards data is more direct than moving data towards figuring", Hadoop uses HDFS to store far reaching data records over the social event. MapReduce gives stream investigating access, runs errands on a get-together of center concentrations, and gives a data controlling structure to an appropriated data stockpiling system. MapReduce number has been used for applications, for instance, making look documents, record gathering, discover the chance to log examination, and diverse specific sorts of data examination. "Make once and read-many" is an approach that licenses data records to be made only once in HDFS and a concise traverse later empowers it to be assessed different conditions over concerning the apportions of doled occupations. In the midst of the structure strategy, Hadoop separates the data into pieces with a predefined square size. The pieces are then made and rehashed in the HDFS. The squares can be imitated various conditions in setting of a specific regard which is set to 3 times obviously. In HDFS, the gathering that Hadoop is gotten is pulled back into two lead pieces, which are (i) the ace concentration called Name Node and (ii) the slaves called Data Nodes. In Hadoop gathering, single Name Node is responsible for general association of the report structure including sparing the information and guiding the occupations to the right Data Nodes that store related application information. Information Nodes bolster Hadoop/MapReduce to manage the occupations with gushing execution in a parallel arranging condition.

ENORMOUS DATA ANALYTICS IN CLOUD ENVIRONMENT

Most corporate ventures confront critical difficulties in completely utilizing their information. Much of the time, information is secured away different databases and preparing frameworks all through the undertaking, and the inquiries clients and examiners solicit require a total view from all information, some of the time totaling many terabytes.

Cerri et al proposed 'Information in the cloud' set up of 'information in the cloud' to bolster communitarian assignments which are computationally concentrated and encourage circulated, heterogeneous learning. This is named as "Utility Computing" gotten from required information all through Cloud the utilities like power, gas for which we pay for what we use from a common asset. With the developing enthusiasm for cloud, investigation is a testing task.In general, Business Intelligence applications, for example, picture preparing, web looks, understanding clients and their purchasing propensities, supply chains and positioning and Bio-informatics (e.g. quality structure expectation) are information concentrated applications. Cloud can be an ideal match for dealing with such expository administrations. For instance, Google's MapReduce can be utilized for investigation as it insightfully pieces the information into littler stockpiling units and appropriates the calculation among minimal effort preparing units. A few research groups have begun taking a shot at making Analytic systems and motors which enable them to give Analytics as a Service. For instance, Zementis propelled the ADAPA prescient investigation choice motor on Amazon EC2, enabling its clients to send, coordinate, and execute factual scoring models like neural systems, bolster vector machine (SVM), choice tree, and different relapse models.

Booz Allen's IT experts, furnished with broad aptitude in the utilization of distributed computing innovation has depicted a route for laying out steps to arrive at acing your enormous information. Cloud innovation joins the prescribed procedures of virtualization, lattice figuring, utility registering, and web advancements. The outcome is an innovation that

https://dx.doi.org/10.22161/ijaers/nctet.2017.9 *ISSN: 2349-6495(P) | 2456-1908(O)*

acquires the readiness of virtualization, the adaptability of framework figuring, and effortlessness of Web 2.0. Distributed computing is a developmental stride in registering that binds together the assets of numerous PCs to work as one substance, permitting the development of greatly adaptable frameworks that can take in and store, prepare and break down the majority of your undertaking's information. The conclusive use of cloud innovation is as an extensive scale information stockpiling, advancement and preparing framework, enabling your endeavor to ace enormous information. In any case, the nimbleness of distributed computing has applications past viable utilization of information. Since all information is currently kept up in a brought together framework, we can help create and execute a unified security arrangement that can be effectively implemented, permitting exact and very much archived control of touchy information. Likewise, the cloud gives a situation in which to model, test, and convey new applications in a small amount of the time and cost of customary frameworks.

The advantages keep on accrueing as your "cloud" develops. As more datasets are totaled, the cloud picks up a minimum amount of information over an endeavor, turning into "the place" to put information. As each dataset is included, and conceivably examined with alternate datasets, there is an exponential increment in advantage to the venture. We can empower your undertaking with streamlined programming and information models, which, consolidated with simple access to an extensive variety of information, results in a blast of development from over your endeavor as information mashups, information mining applications. Couple of decades back, the issue was the lack in data or information. In later past, this issue has been overcome with the appearance of Internet and decreased Storage Memory cost. Be that as it may, another test is the way to break down the information. Information is getting created at a substantially quicker pace than the speed at which it can be prepared with the present framework. Tremendous and devoted servers were produced to tackle this issue. In any case, the issue is with the cost of such a foundation which is not reasonable to every one of the organizations for Availability of information and getting to every last particular reason. So today, these organizations are looking it is the key achievement consider for Cloud figuring which makes possible for every one of these organizations to enlist on an impermanent premise, the computational power and storage room esteem based investigation. Relocating for a particular reason.

MOVING BIG DATA INTO CLOUD

Huge Data is an information investigation approach empowered by late advances in innovations and design. Be that as it may, enormous information involves a colossal duty of equipment and preparing assets, making reception expenses of huge information innovation restrictive to little and medium estimated organizations. Distributed computing offers the guarantee of huge information usage to little and medium estimated organizations.

Huge Data preparing is performed through a programming worldview known as MapReduce. Ordinarily, usage of the MapReduce worldview requires organized appended stockpiling and parallel preparing. The processing needs of MapReduce writing computer programs.

2. RELATED WORK

The grouping arrangement is a fundamental strategy for preparing the data in Bioinformatics, it has an incredible criticalness for finding the capacity and the structure of nucleic acids and protein successions and the data of advancement. This paper quickly portrays the applicable issues of succession arrangement and the most well-known nearby grouping arrangement calculations, Blast calculation. At present, the Blast calculation which given by NCBI or remain solitary can not take care of the real demand for the surge of organic information, this paper accomplishes the Blast-Parallel calculation by further change in view of the Hadoop-Blast calculation. Through serial analyses of the remain solitary Blast calculation and parallelizing tests of the Hadoop-Blast calculation and the Blast-Parallel calculation in view of Hadoop stage, comes about demonstrate that the Blast calculation has altogether higher execution productivity after the parallelization, and the coordinating pace of the Blast-Parallel calculation which has been enhanced can accomplish 1~1.5 times of the Hadoop-Blast algorithm. Today we can create several gig humbles of DNA and RNA sequencing information in seven days for under US$5,000. The shocking rate of information era by these minimal effort, high-throughput innovations in genomics is being coordinated by that of different advances, for example, continuous imaging and mass spectrometry-based stream cytometry.

Accomplishment in the life sciences will rely on

https://dx.doi.org/10.22161/ijaers/nctet.2017.9 *ISSN: 2349-6495(P) | 2456-1908(O)*

upon our capacity to appropriately translate the vast scale, high-dimensional informational collections that are produced by these advances, which thusly obliges us to receive progresses in informatics. Here we talk about how we can ace the distinctive sorts of computational situations that exist —, for example, cloud and heterogeneous figuring — to effectively handle our huge information problems We introduce another way to deal with address the issue of vast succession mining from huge information. The specific issue of intrigue is the compelling mining of long arrangements from extensive scale area information to be functional for Reality Mining applications, which experience the ill effects of a lot of clamor and absence of ground truth. To address this perplexing information, we propose an unsupervised probabilistic theme display called the far off n-gram subject model (DNTM). The DNTM depends on inert Dirichlet distribution (LDA), which is reached out to incorporate successive data. We characterize the generative procedure for the model, determine the derivation methodology, and assess our model on both engineered information and genuine cell phone information. We consider two distinctive cell phone datasets containing normal human portability designs acquired by area detecting, the principal considering GPS/wi-fi areas and the second considering cell tower associations. The DNTM finds significant points on the engineered information and also the two cell phone datasets. At long last, the DNTM is contrasted with LDA by considering log-probability execution on concealed information, demonstrating the prescient energy of the model. The outcomes demonstrate that the DNTM reliably beats LDA as the arrangement length increments.

With the quick improvement of rising applications like relational association examination, semantic Web examination and bioinformatics orchestrate examination, a grouping of data to be readied continues seeing an energetic augmentation. Convincing organization and examination of colossal scale data speaks to an interesting yet essential test. Starting late, huge data has pulled in a lot of thought from the insightful group, industry and also government. This paper displays a couple significant data get ready strategies from system and application viewpoints. In any case, from the point of view of cloud data organization and colossal data get ready frameworks, we demonstrate the key issues of tremendous data taking care of, including circulated processing stage, cloud designing, cloud database and data stockpiling arrangement. Taking after the Map Reduce parallel taking care of framework, we then present Map Reduce change techniques and applications uncovered in the composition. Finally, we discuss the open issues and challenges, and significantly explore the examination course later on immense data planning in disseminated processing circumstances.

We survey the foundation and cutting edge of enormous information. We initially present the general foundation of enormous information and survey related advances, for example, could registering, Internet of Things, server farms, and Hadoop. We then concentrate on the four periods of the esteem chain of huge information, i.e., information era, information obtaining, information stockpiling, and information investigation. For each stage, we present the general foundation, examine the specialized difficulties, and survey the most recent advances. We at long last analyze the few agent uses of huge information, including endeavor administration, Internet of Things, online interpersonal organizations, restorative applications, aggregate knowledge, and brilliant matrix. These talks intend to give an exhaustive review and huge picture to perusers of this energizing zone. This study is finished up with a dialog of open issues and future headings.

3. EXISTING SYSTEM

- In existing Hadoop engineering, NameNode knows the area of the information obstructs in HDFS. NameNode is in charge of allotting the occupations to a customer and partitioning that employment into assignments.

- NameNode additionally appoints the assignments to the TasTrackers (DataNodes). Knowing which DataNode holds the pieces containing the required information, NameNode ought to have the capacity to guide the occupations to the particular DataNodes without experiencing the entire bunch.

- Different contemplates have give some data upgrades and have thought of positive outcomes in view of their presumptions. Others concentrate on the season of instatement and end periods of MapReduce employments

- ShmStreaming presents a Shared memory Streaming mapping to give lockless FIFO line that associates Hadoop and outer projects.

https://dx.doi.org/10.22161/ijaers/nctet.2017.9 *ISSN: 2349-6495(P) | 2456-1908(O)*

DISADVANTAGES OF EXISTING SYSTEM

- The default information appropriation area system causes some poor execution regarding mapping and decreasing undertakings.

- System memory has many issues that could be routed to enhance the framework execution

4. PROPOSED SYSTEM

- In H2Hadoop, before appointing assignments to the DataNodes, we executed a pre-handling stage in the NameNode.

- Our concentrate is on distinguishing and extricating components to manufacture a metadata table that conveys data identified with the area of the information obstructs with these elements. Any employment with similar components ought to just read the information from these particular squares of the bunch without experiencing the entire information once more.

• Proposed Hadoop MapReduce work process (H2Hadoop) is the same as the first Hadoop as far as equipment, system, and hubs. Be that as it may, the product level has been upgraded. We included components in NameNode that enable it to spare particular information in a look into table which named Common Job Blocks Table CJBT.

ADVANTAGES OF PROPOSED SYSTEM

- The proposed arrangement must be utilized for content information.
- BigData, for example, Genomic information and books can be handled effectively utilizing the proposed system.

CJBT stores data about the occupations and the squares related with particular information and elements. This empowers the related occupations to get the outcomes from particular pieces without checking the whole group.

5. MODULES

Hadoop Implementation
BigData Concepts
Native workflow HADOOP H2Hadoop Workflow

MODULES DESCSRIPTION

1. Hadoop Implementation

MapReduce gives stream perusing access, runs undertakings on a group of hubs, and gives an information overseeing framework to a conveyed information stockpiling framework. MapReduce calculation has been utilized for applications, for example, producing look lists, report grouping, get to log investigation, and distinctive different sorts of information examination. "Compose once and read-many" is an approach that grants information documents to be composed just once in HDFS and afterward enables it to be perused many circumstances over concerning the quantities of doled out occupations. Amid the composition procedure, Hadoop isolates the information into squares with a predefined piece estimate

2. BigData Concepts

There are diverse methods for characterizing and contrasting BigData and the customary information, for example, information estimate, substance, gathering and preparing. Huge information has been characterized as extensive informational collections that can't be handled utilizing conventional preparing procedures, for example, Relational Database Management Systems, in a mediocre preparing time. BigData is either a social database (Structured, for example, securities exchange information or non-social database (Semi organized or Unstructured, for example, online networking information or DNA informational collections. The 4V's of BigData are 1) Volume of the information, which implies the information estimate. Some of organizations' information stockpiling is about Zeta byte. Speed, which implies the speed at which the information is produced. Varity of the information, which implies the information frames that diverse applications manage, for example, arrangement information, numeric information or twofold information. Veracity of the information, which implies the vulnerability of the status of the information or how clear the information is to these applications.

3. Native HADOOP Workflow

Framework memory has many issues that could be routed to enhance the framework execution. In Hadoop, Apache plays out a brought together memory approach which is actualized to control the getting the money for and assets. Apache Hadoop bolsters unified information changing. Be that as it

https://dx.doi.org/10.22161/ijaers/nctet.2017.9 *ISSN: 2349-6495(P) | 2456-1908(O)*

may, a few reviews use a disseminated changing way to deal with enhance Hadoop execution. There are distinctive methodologies that examine memory issue. Shm Streaming presents a Shared memory Streaming construction to give lockless FIFO line that associates Hadoop and outer projects.

4. H2HADOOP

In existing Hadoop engineering, NameNode knows the area of the information hinders in HDFS. NameNode is in charge of doling out the occupations to a customer and separating that employment into assignments. NameNode additionally allots the assignments to the Task Trackers (DataNodes). Knowing which DataNode holds the pieces containing the required information, NameNode ought to have the capacity to guide the employments to the particular DataNodes without experiencing the entire group. In H2Hadoop, before allotting errands to the DataNodes, we executed a pre-preparing stage in the NameNode.

6. SYSTEM ARCHITECTURE

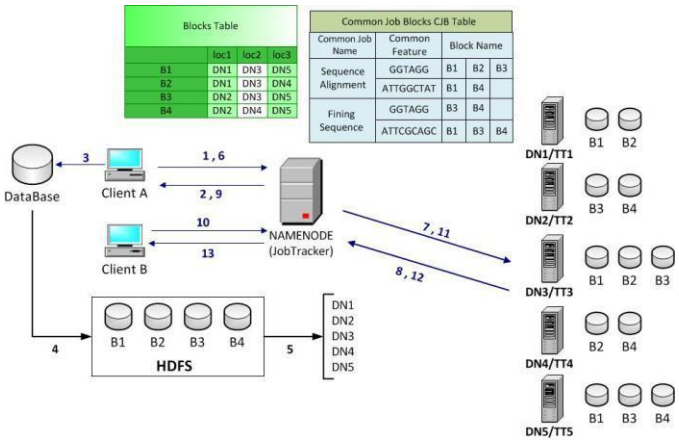

Fig 1: H2Hadoop MapReduce Workflow

Step 1: Client " A" sends a demand to NameNode. The ask for unites the need to duplicate the information records to DataNodes.

Step 2: NameNode replays with the IP address of DataNodes. In the above outline NameNode answers with the IP address of five focus focuses (DN1 to DN5).

Step 3: Client " A" gets to the grungy information for control in Hadoop.

Step 4: Client "A" positions the grungy information into HDFS game plan and fragments squares in light of the information measure. In the above blueprint the pieces B1to B4 are appropriated among the DataNodes.

Step 5: Client "A" sends the three duplicates of every information square to various DataNodes.

Step 6: In this development, customer "A" sends a MapReduce work (job1) to the JobTracker daemon with the source information record name(s).

Step 7: JobTracker sends the assignments to all TaskTrackers holding the squares of the information.

Step 8: Each TaskTracker executes a particular errand on each square and sends the outcomes back to the JobTracker are secured in the CJBT.

Step 9: JobTracker sends the outcome to Client "A". In this development, NameNode keeps the names of the blocks that made the outcomes in the territory request table (CJBT) by the Common Job Name (Job1) that has major part as cleared up as of now.

Step 10: Client "B" sends another MapReduce work "Job2" to the JobTracker with a similar general occupation name and same customary part or super-social occasion of "Job1".

Step 11: JobTracker sends "job2" to TaskTrackers who hold the pieces, which have the essential inevitable result of the MapReduce "Job1" (DN2, DN4, DN5). In this development, the JobTracker begins with checking the CJBT first to discover on the off chance that it is another work which has a tantamount run of the mill name and basic portions of any past ones or not – For this condition yes. By then the JobTracker sends "Job2" just to TT2, TT4 and TT5. We may expect here that the request table will be restored with more reasons for interest OR basically stay as is a result of each time we have another occupation that may pass on a similar name of "Job1".

Step 12: TaskTrackers execute the assignments and send the outcomes back to the JobTracker.
Step 13: JobTracker sends the last outcome to Client "B".

7. CONCLUSION

We present Enhanced Hadoop frame (H2Hadoop), who lets in a NameNode according to discover the blocks into the brush the place secure facts is stored. We mentioned the proposed workflow of H2Hadoop or compared the predicted

https://dx.doi.org/10.22161/ijaers/nctet.2017.9 *ISSN: 2349-6495(P) | 2456-1908(O)*

performance over H2Hadoop in conformity with native Hadoop. In H2hadoop, we examine less data, therefore we bear some Hadoop elements certain so number regarding examine operations, which are reduced by using the quantity over DataNodes assuming the source data blocks, as is recognized above to sending a work to TaskTracker. The most range over statistics blocks so much the TaskTracker pleasure relinquish in imitation of the action is even after the wide variety over blocks as consists of the source data related in accordance with a specific frequent job.

REFERENCES

[1] Ming, M., G. Jing, and C. Jun-jie. *Blast-Parallel: The parallelizing implementation of sequence alignment algorithms based on Hadoop platform*. in *Biomedical Engineering and Informatics (BMEI), 2013 6th International Conference on*. 2013.

[2] Schadt, E.E., et al., *Computational solutions to large-scale data management and analysis.* Nature Reviews Genetics, 2010. **11**(9): p. 647-657.

[3] Farrahi, K. and D. Gatica-Perez, *A probabilistic approach to mining mobile phone data sequences.* Personal Ubiquitous Comput., 2014. **18**(1): p. 223-238.

[4]Marx, V., *Biology: The big challenges of big data.* Nature, 2013. **498**(7453): p. 255-260.

[5]Chen, M., S. Mao, and Y. Liu, *Big Data: A Survey.* Mobile Networks and Applications, 2014. **19**(2): p. 171-209. 9.
Jagadish, H., et al., *Big data and its technical challenges.* Communications of the ACM, 2014. **57**(7): p. 86-94.

[6] White, T., *Hadoop: The definitive guide*. 2012: " O'Reilly Media, Inc.".

[7] Patel, A.B., M. Birla, and U. Nair. *Addressing big data problem using Hadoop and Map Reduce*. in *Engineering (NUiCONE), 2012 Nirma University International Conference on*. 2012.

[8] Buck, J.B., et al. *SciHadoop: Array-based query processing in Hadoop*. in *High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference for*. 2011.

[9] Wu, S., et al. *Query optimization for massively parallel data processing*. in *Proceedings of the 2nd ACM Symposium on Cloud Computing*. 2011. ACM.

[10] Palanisamy, B., et al. *Purlieus: locality-aware resource allocation for MapReduce in a cloud*. in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM.

[11] Matsunaga, A., M. Tsugawa, and J. Fortes. *CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications*. in *eScience, 2008. eScience '08. IEEE Fourth International Conference on*. 2008.

[12] Schatz, M.C., B. Langmead, and S.L. Salzberg, *Cloud computing and the DNA data race.* Nature biotechnology, 2010. **28**(7): p. 691.

[13] Condie, T., et al. *MapReduce Online*. in *NSDI*. 2010.

[14] Herodotou, H., *Hadoop performance models.* arXiv preprint arXiv:1106.0940, 2011.

[15] Lohr, S., *The age of big data.* New York Times, 2012. **11**.

[16] Changqing, J., et al. *Big Data Processing in Cloud Computing Environments*. in *Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on*. 2012.

[17] Hammoud, M. and M.F. Sakr. *Locality-Aware Reduce Task Scheduling for MapReduce*. in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*. 2011.

[18] Dean, J. and S. Ghemawat, *MapReduce: simplified data processing on large clusters.* Communications of the ACM, 2008. **51**(1): p. 107-113.

[19] Li, F., et al., *Distributed data management using MapReduce.* ACM Comput. Surv., 2014. **46**(3): p. 1-42.

[20] Xu, W., W. Luo, and N. Woodward. *Analysis and optimization of data import with hadoop*. IEEE.

[21] Cuff, J.A. and G.J. Barton, *Application of multiple sequence alignment profiles to improve protein secondary structure prediction.* Proteins: Structure, Function, and Bioinformatics, 2000. **40**(3): p. 502-511.

[22] Sadasivam, G.S. and G. Baktavatchalam. *A novel approach to multiple sequence alignment using hadoop data grids*. In *Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud*. 2010. ACM.

[23] Alshammari, H., H. Bajwa, and J. Lee, *Hadoop Based Enhanced Cloud Architecture*, in *ASEE*. 2014: USA.

[24] Erodula, K., C. Bach, and H. Bajwa. *Use of Multi Threaded Asynchronous DNA Sequence Pattern Searching Tool to Identifying Zinc-Finger-Nuclease Binding Sites on the Human Genome*. in *Information Technology: New Generations (ITNG), 2011 Eighth International Conference on*. 2011. IEEE.