Abstract

In the era of internet, people are always connected to each other. People are communicating virtually, irrespective of whether they are living far off or are seated in the same room. Due to the growth of mobile and web applications, internet-enabled devices and services, network traffic is increasing profusely. Various mobile and web-based applications like email, whatsapp, facebook, tweeter, instagram and several other social networking applications are prime sources of huge amount of data in addition to many other transaction processing applications. These sources may produce data which may be text, images, audio or video. Comprehensively, every individual plays a key role in the massive data generation. The pace and scale of data is vast and hence is referred as “Big Data”. Big Data storage, processing, and analytics demand powerful computing infrastructure. Distributed computing is one of the solutions to deal with Big Data. Working with Big Data requires services which are provided by distributed computing such as secured data storage, efficient processing, high resource availability and failure management. There are many tools available, which provide the required services. In recent times, Hadoop is highly in demand due to its ability to store the colossal amount of data and process it on commodity hardware. Hadoop HDFS is widely used for the distributed storage and MapReduce is popular for efficient processing. The data storage using HDFS and data processing using MapReduce consume regular commodity computers, coordinating in form of cluster. Hadoop is a framework used widely in distributed setup, grid setup and even in cloud computing for processing of Big Data. Although Hadoop is widely used, it has huge possibilities of improvisations, for efficient processing, reduce latency, and use of available hardware. Latency in data processing can cause a big loss to giant companies. According to the article presented in January, 2019 in gigaspaces.com by Yoav, Amazon lost 1% of sales in 2009 just because of 100ms of latency. Google also found that 0.5 seconds extra time in search cost them a loss of $4 million per milliseconds. One can imagine how much it is important to optimize performance.

Hadoop has various possibilities of improvements to increase the performance of the cluster. Over time, many researchers have identified several bottlenecks in various phases of Hadoop which decrease the overall performance. The traditional solutions already suggested are application specific, have hardware dependency, involve actual performance variation when implemented in a real cluster, are time-consuming, etc. Moreover, these researchers focussed less on improving heterogeneous cluster performance and to leverage the node processing capability and reduce latency. Therefore, our aim is to optimize the performance of Big Data processing.

To find out the solution we studied Hadoop distributed file system, evaluated the Hadoop schedulers for performance measure. In order to leverage the highly configured heterogeneous systems, we investigate the default HDFS block placement policy. In the current HDFS block placement policy, there is no control on how and where to store data blocks, which in turn, may result in poor load distribution for cluster and degradation of the MapReduce performance. We evaluated that default HDFS block placement policy can be improvised by considering node processing capability for block placement (number of CPU, cores available, and memory available) and heterogeneity of system.

To achieve better performance for big data processing we target upon two important aspects of heterogeneous distributed computing: file system management and process management. First, for file system management we propose “Saksham: the block rearrangement algorithm” which optimizes the processing time by efficient block rearrangement and load balancing in a heterogeneous environment. Our proposed scheme basically controls the block placement and allows us to rearrange the blocks on specified nodes, considering the node’s storage and processing capability. Second, we use the concept of node labelling and scheduling to achieve better process management. In all we propose “Saksham” model which is system independent and leverages the heterogeneity of nodes for performance enhancement.

"Saksham" model has been tested using two benchmark applications and one custom application. The test results prove that the proposed approach has optimized job execution time, reduced latency, data-skew and substantially increased data locality by approximately 25% and decreased the job completion time around 50%.

The overall research work has been published in form of papers as:

1. Padole, M., & Shah, A. (2018). *"Comparative Study of Scheduling Algorithms in Heterogeneous Distributed Computing Systems".* In *Advanced Computing and Communication Technologies* (pp. 111-122). Springer, Singapore.
   [Scopus Indexed, EI-Compendex]
   https://www.springerprofessional.de/en/comparative-study-of-scheduling-algorithms-in-heterogeneous-dist/15161628
2. Shah, A., & Padole, M. (2018). *"Load Balancing through Block Rearrangement Policy for Hadoop Heterogeneous Cluster".* In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 230-236). IEEE.
   [Scopus Indexed, DBLP indexed, EiCompendex and Web of Science THOMSON REUTERS Conference Proceedings Citation Index]
   https://ieeexplore.ieee.org/document/8554404
   https://dblp.org/rec/html/conf/icacci/ShahP18
   https://www.scopus.com/authid/detail.uri?authorId=56716412100
3. Shah, A. and Padole, M., (2019). Performance Analysis of Scheduling Algorithms in Apache Hadoop. In *Data, Engineering and Applications* (pp. 45-57). Springer, Singapore.
   [Book Chapter]
   https://link.springer.com/chapter/10.1007/978-981-13-6351-1_5
4. Ankit Shah, Mamta Padole, (2019) "*Apache Hadoop: A Guide for Cluster Configuration & Testing*", International Journal of Computer Sciences and Engineering, Vol.7, Issue.4, pp.850-855.
   [UGC Approved Journal]
   https://www.ijcseonline.org/pdf_paper_view.php?paper_id=4118&135-IJCSE-06624.pdf

5. *"Hadoop Performance Acceleration by Effective Data and Job Placement".*
   Paper submitted for Advances in Intelligent Systems and Computing, Springer, Singapore
   [Accepted] [Scopus Indexed, EI-Compendex]
6. *"Saksham: Resource Aware Block Rearrangement Algorithm for Load Balancing in Hadoop".* Paper submitted in IGI Global book chapter on "Big Data Analytics for Sustainable Computing"