A REVIEW ON HIGH PERFORMANCE COMPUTING FOR SATELLITE IMAGE USING DISTRIBUTED COMPUTING WITH APACHE SPARK

Pallavi U. Hiwarkar [1] , Mangala S. Madankar [2] , T P Girish Kumar [3]

(1 PG Scholar , 2 Assistant Professor, 3 Scientist/Engineer 'SE')

[1,2] Department of Computer Science And Engineering, G. H. Raisoni College of Engineering,

[3]Regional Remote Sensing Centre-Central, RRSC.

Abstract : High Performance Computing (HPC) is that the recently developed technology at intervals the sector of technology, that evolved owing to meet increasing demands for[17] method speed and analysing/processing massive size of data sets. HPC brings along many technologies like laptop design, algorithm, programs to solve/handle advanced complicated issues quickly and effectively. This technology focuses on developing and implementing methods like distributed computing for solving problems[18]. Distributed computing is that the method of aggregating the facility of many computing entities to collaboratively run one procedure task in a very clear and coherent means, in order that they seem as one centralized system Connecting users and resources in a very clear, open and ascendible means is that the main goal of distributed software package. We do believe that a single system would be unable to handle the processing large data. Thus, comes the need for Distributed Systems. Satellite image process plays a significant role for the analysis developments in natural philosophy, Remote Sensing, GIS, Agriculture observance, Disaster Management[19] and plenty of alternative fields of study. However, process those satellite pictures need an outsized quantity of computation time because of its complicated and huge process criteria. This looks a barrier for real time deciding to modify the work quicker, distributed computing is an suitable solution.in this project we also use apache spark, the apache spark is a lightning-fast cluster computing designed for fast computation. Python will be used for programming[20]. Batch processing of Satellite imagery takes more time. By using parallel processing we can improve the performance.

Keywords: Satellite images; Apache spark; python; Image processing.

I. INTRODUCTION

Remote sensing is one of the domains where the data have shown an explosive growth[21]. As the higher resolution observations are increasing the volume of the remote sensing imagery is also increasing at a higher rate. The RS data being generated from a single satellite on a daily basis is in terabytes, which makes it important to process it proficiently. In recent years, focusing on massive data storing and processing, high-speed data flows, multisource, the remote sensing image information processing technologies developed rapidly [1]. Traditional PCs never satisfy the real-time processing demanding. In order to largely increase the computing speed, processing performance and execution efficiency, multi-core technologies and parallel computing theories were adopted.

High Performance Computing (HPC) is that the recently developed technology amoung the sector of computing science, which evolved because of meet increasing demands[22] for processing speed. HPC brings along many technologies like computer architecture, algorithm, programs and computer code below one cover to resolve advanced issues quickly and effectively. This technology focuses on developing and implementing methods like distributed processing for solving problems. Satellite image process plays a significant role for the analysis developments in natural philosophy, Remote Sensing, GIS, Agriculture observance, Disaster Management[19] and plenty of alternative fields of study. However, process those satellite pictures need an outsized quantity of computation time because of its complicated and huge process criteria. This looks a barrier for real time deciding to modify the work quicker, distributed computing is an suitable solution and also distributed system contains multiple nodes that are physically separate however coupled along mistreatment the network[23]. All the nodes during this system communicate with one another and handle processes in wheel.

Spark could be a project of the Apache code Foundation, originally created by the AMP research laboratory at the University of Calif, Berkeley, regarding the performances .Spark adopts the MapReduce paradigm, and it's accessible by exploitation the API by completely different programming languages (such as Scala, Java, and Python)[24].

Application of HPC technology is obtaining additional importance in remote sensing analysis work[25]. The employment of HPC systems in remote sensing applications has become additional and additional widespread in recent years HPC is in position to enhance the computing[26] speed to a good extent in huge processing , that makes itself an efficient thanks to solve the matter of process potency in remote sensing information. During this paper we have a tendency to gift techniques and ways of High Performance Computing for remotely perceived satellite image process and analyzing. The subsequent sections in short describe the High Performance Computing technology for remote sensing processing and analyzing ways.In the remote sensing field, many researchers have been using parallel computing techniques to accelerate clustering for RSBD (Remote Sensing Big Data)[27].

II. DISTRIBUTED COMPUTING

Distributed computing is that the method of aggregating the facility of many computing entities to collaboratively run one procedure task in a very clear and coherent means, in order that they seem as one centralized system. Connecting users and resources in a very clear, open and ascendible means is that the main goal of distributed software package[30]. Godfrey B.(2002) has delineate that distributed computing works by cacophonous the larger into smaller chunks which might be performed at an equivalent time severally of every alternative same [31]. The two main entities in distributing computing square measure the server and therefore the many purchasers .A central computer, the server can generate work packages that square measure passed onto employee purchasers . The purchasers can perform the task, careful in a very work package knowledge and once it's finished the completed work package are passed back to the server. The operating method of semi distributed programming policy is given within the Figure 3.

Process image knowledge generated by new remote sensing systems will severely tax the procedure limits of the classic single processor systems that square measure[28] usually offered to the remote sensing professional person[29].in operation on these giant knowledge sets with one ADP system, typically simplifying approximations square measure used which will limit the exactitude of the ultimate result. Recent work geographical region National Laboratory powerfully counsel strongly suggest that a distributed network of cheap PCs is designed that's best to upset intensive computationally issues. The new kind of distributed computing can take away procedure constraints; image process algorithms for remote detected pictures square measure currently being thought of .

Geo referencing is basic perform of remote sensing data processing. It's a method of assignment geographic information to an image[18].

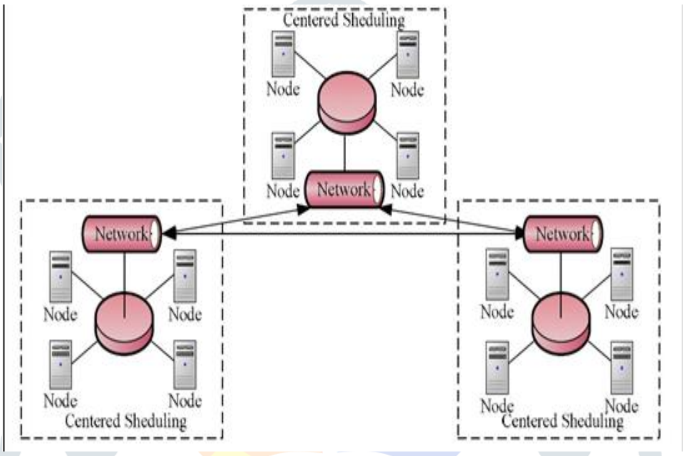

Figure 3. Semi distributed scheduling policy (modified after, Hangye Liu et al., 2009)

2.1 ADVANTAGES OF DISTRIBUTED SYSTEMS

- All the nodes within the distributed system are connected to every alternative. thus nodes will simply share data with alternative nodes.
- More nodes will simply be additional to the distributed system i.e. it are often scaled as needed.
- Failure of one node does not cause the failure of the complete distributed system. alternative nodes will still communicate with one another[30].

III. APACHE SPARK

Apache Spark is thought as a quick, easy-to-use and general engine for giant processing that has Intrinsical modules for streaming, SQL ,Machine Learning (ML)[31] an graph process. Apache Spark may be a leading, ASCII text file cluster computing computer code framework fashionable Programmer sand information scientists operating with huge information. Its economic programs will run the maximum amount as one hundred times quicker than Hadoop Map Reduce jobs. Spark facilitates the implementation of every unvaried algorithms that visit their information set multiple times during a loop, and interactive/exploratory data analysis, i.e., the repeated database-style querying of data Apache Spark achieves high performance for each batch and streaming information, employing a progressive DAG computer hardware, a query optimizer, and a physical execution engine. Spark offers over eighty high-level operators that create it straight forward to create parallel apps. And you'll be able to use it interactively from the Scala, Python, R, and SQL shells[32].

In our experiments we are using spark with python, Spark has some in-memory processing capabilities, although it doesn't store all the data in the memory. The core conception of Apache Spark that has helped Spark in achieving high performance is Resilient Distributed Datasets (RDD).RDD can be persisted in the memory as well on the disk. Keeping the RDD within the memory makes the method quicker because the same RDD are often used multiple times while not computing it time and once more .RDDs are meant to be designed for repetitious algorithms and execution that applies identical operation on all the records of the information set. Jeremy citizen has place concentrate on Apache Spark whereas discussing the open supply tools for giant scale neurobiology [15].

Spark is not a modified version of Hadoop and is not, really, dependent on Hadoop because it has its own cluster management[33]. Hadoop is just one of the ways to implement Spark.

3.1. FEATURES OF APACHE SPARK

Apache Spark has following features.

- **Speed** − Spark helps to run an application in Hadoop cluster, up to a hundred times quicker in memory, and ten times quicker once running on disk. This is often realizable by reducing form of read/write operations to disk.
- **Supports multiple languages** − Spark provides intrinsic genus APIs in Java, Scala or Python. Therefore, you'll be able to write applications in several languages.
- **Advanced Analytics** − Spark not exclusively supports 'Map' mand 'reduce'. It collectively supports SQL queries, Streaming knowledge, Machine learning (ML), and Graph algorithms.

3.2 PART OF SPARK

Spark is not a modified version of Hadoop and is not, really, dependent on Hadoop because it has its own cluster management. Apache Spark is thought as a quick, easy-to-use and general engine for giant processing that has Intrinsical modules for streaming, SQL ,Machine Learning (ML) an graph process. Hadoop is just one of the ways to implement Spark. The following illustration depicts the different part of Spark.

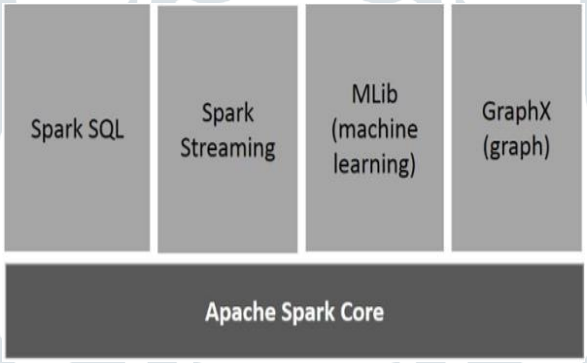

Figure 2. Part of spark

- **Apache Spark Core:-**Spark Core is that the underlying general execution engine for spark platform that every one various usefulness is built alternative upon.It provides In-Memory computing and referencing datasets in memory device systems.
- **Spark SQL:-**Spark SQL might be a district on high of Spark Core that introduces a current information abstraction referred to as Schema RDD,that provides support for structured and semi-structured information.
- **Spark Streaming:-**Spark Streaming leverages Spark Core's fast programing capability to perform streaming analytics. It ingests information in mini-batches and performs RDD (Resilient Distributed Datasets) transformations on those mini-batches of knowledge.
- **MLlib (Machine Learning Library):-**MLlib might be a distributed machine learning framework on high of Spark owing to the distributed memory-based Spark style. It is, in line with benchmarks, done by the MLlib developers against the Alternating Least Squares (ALS) implementations. Spark MLlib is ninefold as fast as a result of the Hadoop disk-based version of Apache driver (before driver gained a Spark interface).
- **GraphX:-** GraphX might be a distributed graph-processing framework on high of Spark. It provides associate in nursing API for expressing graph computation which is able to model the user-defined graphs by exploitation Pregel abstraction API. It collectively provides associate in nursing optimized runtime for this abstraction.

3.3. SPARK ARCHITECTURE OVERVIEW

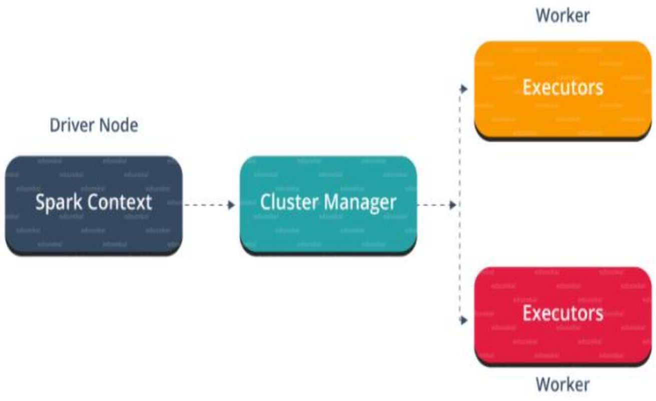

Figure 3. Architecture of spark

Spark Applications contains a driver method and a collection of fiduciary processes. The driver process runs your main() function, sits on a node with in the cluster, and is accountable for three things: maintaining data concerting the Spark Application; responding to a user’s program or input; and analyzing, distributing, and planing work across the executors (defined momentarily).

The driver method is completely essential - it’s the guts of a Spark Application and maintains all relevant information throughout the period of time of the application.The executors square measure chargeable for really corporal punishment the work that the driving force assigns them. This means, every fiduciary is changeable for solely 2 things: executing code assigned to it by the driver and reporting the state of the computation, on that executor, back to the driver node. The cluster manager controls physical machineries and provide resources to Spark Applications.

4. LITERATURE SURVEY

In the previous section, we tend to mentioned some basic information concerning distributed computing and therefore the want of developing an answer to method them in Bigdata surroundings. In recent years, there are multiple studies around the processing of remote sensing pictures on Hadoop. Wangetal planned the answer for enormous pixellevel remote sensing image processing on Hadoop [3]. In [4], researchers show however parallel implementation of Kauffman’s initialization will improve the performance and quantifiability. Kmeans formula suffers from initial cluster bigining points which might be improved victimisation Kauffman’s initialization. however the latter formula is slower and thus enforced on Hadoop MapReduce. Sharma et al. planned an answer whereby the satellite image agglomeration is performed in a very distributed manner for multiple values of k in a very K-means++ formula [5]. The paper conjointly showed however Silhouette Index may be enforced on the pictures to understand the suitable variety of teams in a picture. In [6] conjointly, the paper suggests to run K-means agglomeration of satellite pictures on Hadoop and showed the importance of MapReduce.

Scientists and researchers have conjointly worked on image retrieval victimisation Hadoop. Luca et.al planned an answer to mix each Apache hadoop and Apache Spark for retrieval for an enormous assortment [7]. The creation of indexes was performed victimisation MapReduce whereas the retrieval was dead victimisation Spark. Apache Spark is another distributed framework just like MapReduce [8]. In a number of the things Spark processes quicker than MapReduce thanks to its in-memory processing mechanism. Wei et al demonstrates the aptitude of Spark by process remotely perceived information in-memory [9]. Noha et.al introduces a unique approach of Bag of Visual Words supported Chain-clustering binary search tree to be enforced on MapReduce for content based mostly image retrieval [10].

Mohammad H. Almeer conducted the experiments to indicate the quantifiability of process of pictures on MapReduce. The paper clearly showed the numerous within the time interval once the process happens on one pc as compare to MapReduce [11]. The study conferred the methodology of land cowl recognition resolution supported Hadoop. A scalable modeling system was enforced in MapReduce framework [12].

In [13], researchers planned the K-Means agglomeration formula that run in parallel supported MapReduce. Zhenhua et al. in [14] ran the MapReduce K-means agglomeration on satellite pictures. Taking the assistance of MapReduce execution framework, the formula scaled as good as on artifact hardware.

Later, [1] came up with ascendable K-means++ to optimize it additional by sampling additional purposes in every iteration rather than single point. Despite the fact that the amount of iteration decreases still several iterations are required. In [6], the paper planned associate degree economical k-means approximation with MapReduce. It planned that one MapReduce is enough for the initialisation part rather than multiple iterations. Just in case of MapReduce, multiple iterations of MapReduce jobs are required to seek out the suitable variety of clusters. To beat the respetitive MapReduce approach, Garcia and Naldi planned running identical MapReduce jobs for all the values of k [15]. The ultimate output of the MR jobs consists of set of clusters for all the values of k. They used the Simplified Silhouette index analysis to decide on the proper price of k. They compared the execution results with Apache driver for information sets of various sizes generated by the MixSim R package. The review shows the parallelizing iteration phase for generating clusters. The paper failed to conduct the experiment on real information and therefore the initialisation part wasn’t parallelized.

Satellite image process plays an important role for analysis developments in natural philosophy, Remote Sensing, GIS, Agriculture observation, Disaster Management and plenty of different fields of study. However, process those satellite pictures need an oversized quantity of computation time because of its advanced and enormous process criteria. This appears a barrier for real time deciding to modify the duty quicker, distributed computing is an appropriate resolution. Recently, Cluster and Grid square measure

2 most acquainted and powerful distributed systems to serve for top performance parallel applications. GRASS GIS (Geographical Resources Analysis Support System) is an open source software/tool, that has been wont to method the satellite pictures. within GRASS, completely different modules are developed for process satellite images. GRASS module “r.vi” is developed by Kamble and Chemin, and is employed as a check example for this study. Developing the methodology, that permits to run GRASS GIS surroundings for satellite pictures process on distributed computing systems, is that the main regarding issue of this paper[16].

5. OVERVIEW OF TOPIC

Distributed computing technique is one amongst the set of High Performance Computing (HPC) techniques that are being employed today to process large amounts of data. Distributed processing is especially helpful in comes that need complicated computations. HPC brings along many technologies like laptop design, algorithm, programs to solve/handle advanced complicated issues quickly and effectively. Satellite image process plays a significant role for the analysis developments in natural philosophy, Remote Sensing, GIS, Agriculture observance, Disaster Management and plenty of alternative fields of study. However, process those satellite pictures need an outsized quantity of computation time because of its complicated and huge process criteria. This looks a barrier for real time deciding to modify the work quicker, distributed computing is an suitable solution and also distributed system contains multiple nodes that are physically separate however coupled along mistreatment the network. All the nodes during this system communicate with one another and handle processes in wheel. the apache spark could be a lightning-fast cluster computing designed for quick computation. Although, a lot of analysis work needs on satellite data processing associate degreed analyzing over HPC platform for obtaining an increased and quick output for varied remote sensing applications. In our experiments we are using spark with python, Spark has some in-memory processing capabilities, although it doesn’t store all the data in the memory.

6. CONCLUSION

Distributed computing technique is one amongst the set of High Performance Computing (HPC) techniques that are being employed today to process large amounts of data. data In our experiments we have improve the speed of large amount of satellite data we are using distributed computing and apache spark for fast computation we are using Landsat 7 images for fast processing . We do believe that a single system would be unable to handle the processing large data. Thus, comes the need for Distributed Systems. Satellite image process plays a significant role for the analysis developments in natural philosophy, Remote Sensing, GIS, Agriculture observance, Disaster Management and plenty of alternative fields of study. However, process those satellite pictures need an outsized quantity of computation time because of its complicated and huge process criteria. This looks a barrier for real time deciding to modify the work quicker, distributed computing is an suitable solution In our experiments we are using spark with python, Spark has some in-memory processing capabilities, although it doesn’t store all the data in the memory. Application of HPC technology is getting more importance in remote sensing research work. The utilization of HPC systems in remote sensing applications has become more and more widespread in recent years HPC is able to improve the computing speed to a great extent in massive data processing. more research work requires on satellite data processing and analyzing over HPC platform for getting an enhanced and fast output for various remote sensing applications.

REFERENCES

[1] **B. Bahmani, B. Moseley, A. Vattani, R. Kumar and S. Vassilvitskii, “Scalable k-means++,” in International Conference on Very Large Databases, 2012.**

[2] **Apache, "Apache Hadoop," [Online]. Available: http://hadoop.apache.org/. [Accessed 28 3 2017].**

[3] **X. Wang, G. Li, W. Yu and Q. Zou, "Research on Method for Massive Pixel-Level Remote Sensing Image Processing Based on Hadoop," *Applied Mechanics and Materials,* Vols. 333-335, pp. 1224-1230, 2013.**

[4] **H. Xia, H. A. Karimi and L. Meng, "Parallel implementation of Kaufman's initialization for clustering large remote sensing images on clouds," *Computers, Environment and Urban Systems,* 2014.**

[5] **T. Sharma, V. Shokeen and S. Mathur, "Multiple K Means ++ Clustering of Satellite Image Using Hadoop MapReduce and Spark," *International journal of advanced studies in computer science and engineering, vol. 5, no. 4, 2016.***

[6] **Z. Lv, Y. Hu, Z. Haidong, J. Wu, B. Li and H. Zhao, "Parallel K-Means Clustering of Remote Sensing Images Based on MapReduce," in *2010 International Conference on Web Information Systems and Mining, WISM 2010*, 2010.**

[7] **L. Costantini and R. Fondazione, "Performances Evaluation of a Novel Hadoop and Spark Based System of Image Retrieval for Huge Collections," *Advances in Multimedia,* 2015.**

[8] **"Apache Spark," [Online]. Available: http://spark.apache.org/.**

[9] **W. Huang, L. Meng and D. Zhang, "In-Memory Parallel Processing of Massive Remotely Sensed Data Using an Apache Spark on Hadoop YARN Model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* pp. 3-19, 2016.**

[10] **N. A. Sakr, A. I. ELdesouky and H. Arafat, "An efficient fast-response content-based image retrieval framework for big data," *Computers and Electrical Engineering,* 2016.**

[11] **M. H. Almeer, "Hadoop Mapreduce for Remote Sensing Image Analysis Mohamed," *International Journal of Emerging Technology and Advanced Engineering,* 2012.**

[12] **N. . C. F. Codella, G. Hua, A. Natsev and J. R. Smith, "Towards large scale land-cover recognition of satellite images," in *ICICS 2011 - 8th International Conference on Information, Communications and Signal Processing*, 2011.**

[13] W. Zhao, H. Ma and Q. He, “Parallel K-Means Clustering Based on MapReduce,” in CloudCom, Beijing, 2009.

[14] Z. Lv, Y. Hu, Z. Haidong, J. Wu, B. Li and H. Zhao, "Parallel K-Means Clustering of Remote Sensing Images Based on MapReduce," *in 2010 International Conference on Web Information Systems and Mining, 2010*.

[15] K. D. Garcia and M. C. Naldi, "Multiple Parallel MapReduce k-means Clustering with Validation and Selection," *in Brazilian Conference on Intelligent Systems, IEEE, 2014.*

[16] Shamim Akhter, Yann Chemin, Kento Aida in Tokyo Institute of Technology Asian Institute of Technology "Satellite Image Processing on Distributed Computing Environments*" IEEE,2009.*

[17]https://www.researchgate.net/topic/Satellite-Image-Processing/publications

[18]https://www.techopedia.com/definition/4595/high-performance-computing-hpc, https://www.slideshare.net/journalsats/ijcatr02041007

[19]https://www.researchgate.net/scientificcontributions/48816711_P_Thangaraj

[20]https://www.tutorialspoint.com/apache_spark/apache_spark_tutorial.pdf

[21]https://www.sciencedirect.com/science/article/pii/B9780128124437000077

[22]https://pdfs.semanticscholar.org/d23c/8cce17edf443965ffab62e55638d110df313.pdf

[23]https://www.int-arch-photogramm-remote-sensspatial-inf-sci.net/XLII-2-W13/909/2019/isprs-archives-XLII-2-W13-909-2019.pdf

[24] https://www.hindawi.com/journals/am/2015/629783/

[25]https://www.researchgate.net/publication/261210310_Remote_sensing_data_fusion_algorithms_with_parallel_computing

[26]https://www.researchgate.net/publication/301915899_Approximate_Computing_of_Remotely_Sensed_Data_SVM_Hyperspectral_Image_Classification_as_a_Case_Study

[27] https://www.mdpi.com/1424-8220/19/15/3438/htm

[28]https://www.researchgate.net/publication/215608232_Grid_Service_Based_on_GIMP_for_Processing_Remote_Sensing_Images

[29] https://www.mdpi.com/journal/remotesensing

[30]https://docs.oracle.com/cd/A58617_01/server.804/a58238/ch1_unde.htm

[31]https://www.enterprisesearchblog.com/machine-learning/

[32] https://blog.jetbrains.com/pycharm/2019/02/pycharm-2019-1-eap-3/

[33]https://oracle151.wordpress.com/2016/02/13/best-apache-spark-training-institute-in-chennai-2