IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

A New Approach in Grid Computing using Procedure distribution for high performance computing

Maitry Joshi[1*], Ankit Shah[2]

[1]Dept.Computer Engineering, Shankersinh Vaghela Bapu Institute of Technology, Vasan, Gujarat, India

Corresponding Author: maitry.er@gmail.com, Tel.: +91-9426246651

DOI: https://doi.org/10.26438/ijcse/v7i5.8691 | Available online at: www.ijcseonline.org

Accepted: 12/May/2019, Published: 31/May/2019

Abstract— With the current advances of today’s technology in many sectors like manufacturing, business and web application such as Variety of data to be processed continues to witness an exponential rise. To utilize the numerous benefits of grid computing, Hadoop, HPC techniques should be integrated in the current grid environment. In this paper, the definition features and requirements of distribute process should be distributed to techniques use Hadoop is suggested as it the most commonly used technique in handling process distributed as it offers reliability, ease to use and ease to maintenance and scalability. High Performance Computing (HPC) uses to distribute computational cycles of searching or Time and process jobs and decrease the amount of time in a single job would take. A HPC processing jobs typically consist of searching a time and process the jobs. The process is divided in the form of Grid using Grid Computing. This new approach analyzes the process, distributed among grid and decrease job run time, so to produce the optimized result.

Keywords— Grid computing; Hadoop; High Performance Computing; Distributed computing; Hadoop Distributed File System (HDFS).

I. INTRODUCTION

Grid computing main different things that different individuals, In grid-computing can capture to ones imagination and indeed some days to be come to pass in an actualism there to be many things of technical, business, social and political an point at issue that need to be located, If we are considering to these kind of vision and an ultimate aim, there are soo numerous little steps that to necessary to be taken into achieve it [1]. A smaller steps can profit to getting an own aim. A Grid computing history of an along path of different solution & technologies and they moves up to nearest and final aim. It's major value is an underlying the distributed computing infrastructure technologies and they are involving supports of cross-organizational resource sharing, in a words, an application and virtualization across the platforms, technologies, and organizations. Grids have come into view as a cyber-supporting structure for the coming newly generation of commercial business and Science an applications and many technologies to be supported and grids used an integrating distributed large Scale and heterogeneous resources. A Gridsim middleware have a multiple tool kit list like an “Alchemi, globus, legion, Gridbus & UNICORE and many more…” have been done and created in order to manage an supporting structure that enables to transparently or communicate as securely to the users to the access remote resources to be allot to world wide area network. A Workflow is related to an automatic process, Which data or files can passed between the participants in the opinion of the defined set of rules and their order to be achieved aim. A system of workflow management to be defined to manage and execute task which given by workflow on the resources.

See the below figure-1.1 to shows the system of workflow management architecture of grid computing. Generally in system of workflow can be created a user to using modeling tools or to be created automatic aid of grid data or guidance to services like MDS-Monitoring & discovery services and VDS-virtual data system to in priority in run time.

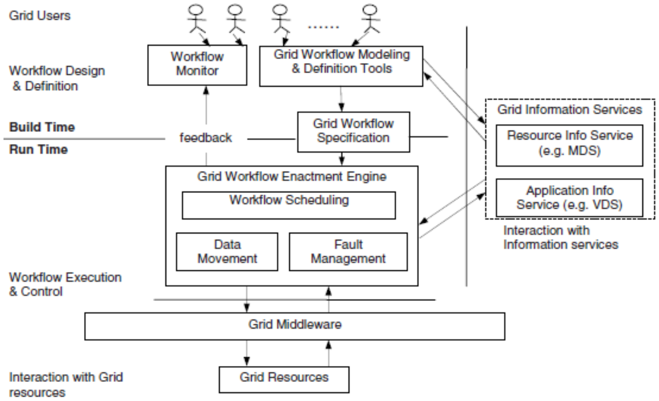

Grid Users
......
Workflow Design & Definition
Workflow Monitor
Grid Workflow Modeling & Definition Tools
Grid Information Services
Build Time
Run Time
feedback
Grid Workflow Specification
Resource Info Service (e.g. MDS)
Grid Workflow Enactment Engine
Workflow Scheduling
Application Info Service (e.g. VDS)
Data Movement
Fault Management
Interaction with Information services
Workflow Execution & Control
Grid Middleware
Interaction with Grid resources
Grid Resources

Fig. 1.1. Grid Workflow Management System

A workflow specification is defined to their task of process or an activity and control the all data deepen activities or tasks and their control of depended data. A workflow system can be run at time then legislation of engine to be manage to the execution in system by utilization in this middleware. A workflow management system have a three major components are be involved. A first component is workflow scheduling, second one is data movement and last one is fault management, these scheduling resources and allocated task is suitable resources is used to users as per users requirement. While the execution process the system is failure during the process then faultmanagment mechanism handle. A faultmanagment mechanism to manage the all data movements and the transferred data in resources In addition of this system should be provide a feedback or an review to the monitor for the client or user to view their system process and check their status through in grid workflow monitor.

- GridSim a software platform that enables users to model and simulate the characteristics of Grid resources and networks with different configurations.
- A large datasets on the cluster of commodity hardware in distributed processing should be allowded by the open source framework.
- Hadoop is a data Management tool and uses scale out storage and processing.

Developing the data processing an application should be executed in distributed computing environs in open source S/W framework hadoop.

An Applications are built into using a HADOOP'S run in to large data sets distributed an across the clusters of commodity hardware, A Commodity computers are cheeps and more available, these are mainly useful for an achieving greater computational power in low cost.

Hadoop is an S/W framework; it does can be installed on a commodity an operating Linux cluster to permit the large scale of distributed data analysis. A Hadoop is provides the Java-based API as well as Hadoop Distributed File System (HDFS) that allows to parallel processing an across the nodes of cluster. In HDFS-Hadoop distributed file system files should be divided into the multiple blocks

Files in HDFS-Hadoop distributed file system are divided into the multiple blocks and replicat into other Data Nodes. These files divided by defaults in two nodes into ascertain high data availability and durability in those cases whose are failer during the execution process or job like parallel processing in hadoop environment. Fundamentally hadoop cluster have two types of node to be operand. See the first one is master and second is salve or same like master and worker pattern, In hadoop distriuted file system NameNode is a master node and DataNodes are the workers node. In slave node like worker node means a Datanodes have actual files part on these node can stored. Then the master node like Namenode containes some information can be stored in different file blocks to be located, when a system to be starts block to many changes one DataNode to be another DataNode but, when it's reported to the client or Namenode-Slave node who's submitted their job (mapreduce) or own data periodically. A block resides and communicates their data nodes files to all these information is gated by client. NameNode include only metadata. The existing technology such as grid has access to huge amounts of computing power by summing of resources and offering a single system view, These technologies have become an influential architecture that performs computing, An addition of an importance aim of these technologies ate to the convey computing answer to solving large amount of data issues, such as multi-media, large scale, and high dimensional data sets.

II. BACKGROUND & RELATED WORK

"A hadoop cluster is a collection of independent components connected through a dedicated network to work as a single centralized data processing resource. "

"A hadoop cluster can be referred to as a computational computer cluster for storing and analysing big data (structured, semi-structured and unstructured) in a distributed environment."

"A computational computer cluster that distributes data analysis workload across various cluster nodes that work collectively to process the data in parallel."

A hadoop clusters are known that "Shared Nothing" systems becose it's nothing is allocated between the "nodes in a hadoop cluster except for the network which connects them". An Allocate is nothing but paradigm of an "hadoop cluster reduces the processing latency" soo, when there is a need to process queries on "huge amounts of data the cluster-wide latency is completely minimized".

1. Paper: Enhancing dataset Processing in hadoop YARN performance for big data application
Author: Ahmed.A. A., Dae-Ki K. & M. Kim
Publication: Springer
Summary: In these paper hadoop "MapReduce" file shoul be distributed, In system input datafiles should be fully loaded and distributed to worker, A worker start the computational work according to the user logic and their needs altercation of resource between reduce and map to contention of resources between the map stages notable to extended their execution time specially in memory IO overheads. In these paper to optimize the present null schedule slots to the local memory management and get the minimum execution time.

2. Paper: Big Data Analysis Using Hadoop Cluster
Author: Saldhi Ankita, Goel Abhinav, Yadav Dipesh, Saldhi Ankur, Saksena Dhruv, S. Indu.
Publication: IEEE

Summary: This paper proposes to execute tasks assigned to a single DataIn these paper process should be execute their task or process for single data node is sequentially, in these paper to be propose a bunch for single data node for SMs-Streaming multi-processors. A large amount of data or process should be come to the different different sources to run on parallelly in a hadoop cluster and get the result efficiently should be applied these methodology for industrial or business.

3. Paper: An Analytical performance model of mapreduce
Author: Xiao.Y, Jianling .S
Publication: IEEE
Summary: In these paper propose a understanding and general MapReduce performance model for the excellent for their components for comprehensive program performance & checked it into a smaller cluster, in these paper to get solution to "indicate there model can predict a perform of MapReduce system &it's relation" to be config.

4. Paper: A Brief Preview of Efficient Hadoop Job Schedulers
Author: Mukesh Singla
Publication: Research India Publications
Summary: This paper helps in brief examination of various schedulers for efficient data. In this paper we have studied many techniques for making the efficient scheduler so that we can speed up our system or data retrieval. Different schedulers like Dynamic Hadoop Fair Schedulers improves the improves the performance and utilization of the Hadoop cluster, Octopus, develops a multi-job fair scheduler by considering the node capabilities, Real-Time MapReduce (RTMR) scheduler provides better cluster utilization and ratio of job success, FSPY (Fair Sojourn Protocol in YARN) scheduler improve responsiveness with guaranteeing fairness by calculating job virtual sizes etc.

5. **Paper:**Big Data Implementation Using hadoop and grid Computing
Author: Yadav Pratik A., Sase Yuvraj S.
Publication:IJIRSET
Summary: In these paper is focused only those methods in which to be used a grid-computing almong with hadoop, AGrid-Computing should be provide a largest "storage capability and computation power". These paper listed some "open source toolkit use and an implementation solution like:Hadoop, Globus Toolkit".

6. **Paper:**Implementing big data management on Grid-Computing environment.
Author: Lawal M. A.
Publication:IJECS
Summary: In these paper to be represent to the data referd to a big data, then a big data to be manage & process on the "data pose on interesting nut significant problem", during the mainly advantage of grid computing-big data process & manage the technique to be integrated, in these paper to be defined their feature & their needs among the big data-platform.

7. **Paper:**A Big Data implementation based on Grid Computing.
Author: Garlasu Dan;Sandulescu Virginia; Mariana.M.
Publication: IJECS
Summary: In these paper to be main purpose of these "article should be represent to the present a wayof processing big data using Grid-Technologies" -For that;framework for managing a "Big Data will be presented along with the way to implement it around a grid architecture".

8. **Paper:**A Study on hadoop_MapReduce_Techniques and Applications on Grid.
Author: Savant Ila;Muke Richa; Narlawar Nilay.
Publication: IJERT
Summary: In thess paper focused on hadoop's_MapReduce-techniques & there about study, in these paper to be discussed on MapReduce-Application on "grid-computing, image processing to deal with big data" isseues.

9. **Paper:**Big data accomplishment using apache hadoop & grid computing environment.
Author: Dr.B.venkata R.Reddy;G.Narendra; R.Navateja R.
Publication:IJIRCCE
Summary: In these paper to some lists of the "open source toolkits to accomplish the solutions such as Hadoop, Globus Toolkit"Grid-Computing to be "provide a huge amount of storage capability in distributed manner with high computation power".

III. PARAMETERS USED

1) **Load Balancing:** Load balancing component balances job distribution among cluster nodes. In Ignite load balancing is achieved via LoadBalancingSpi which controls load on all nodes and makes sure that every node in the cluster is equally loaded. In homogeneous environments with homogeneous tasks load balancing is achieved by random or round-robin policies. However, in many other use cases, especially under uneven load, more complex adaptive load-balancing policies may be needed.

2) **Latency Decrease:** A low latency define to a computer n/w that is "optimized to process a very high volume of data messages with minimal delay or latency". This n/w is designed to help of the operands those are needed require for near "real-time access to rapidly changing data".

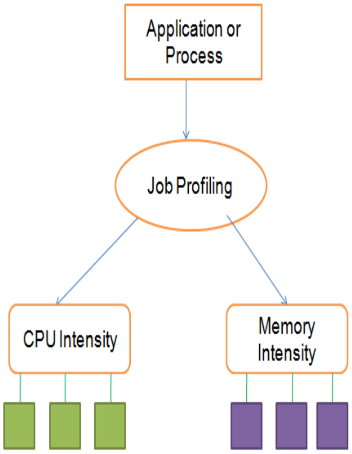
3) **Data locality:** In hadoop, Data locality is the process of moving the computation close to where the actual data resides on the node, instead of moving large data to computation. This minimizes network congestion and increases the overall throughput of the system. This feature of Hadoop we will discuss in detail in this tutorial. We will learn **what is data locality in Hadoop**, data locality definition, how Hadoop exploits Data Locality, what is the need of Hadoop Data Locality, various types of data locality in Hadoop MapReduce, Data locality optimization in Hadoop and various advantages of Hadoop data locality.

4) **High Performance:** HPC clusters will typically have a large number of computers (often called 'nodes') and, in general, most of these nodes would be configured identically. Though from the outside the cluster may look like a single system, the internal workings to make this happen can be quite complex.

IV. PROPOSED WORK

There are lots of researches done on various Hadoop Cluster related to distribute the process. They try to improve the client side request to application or task to be distributed and getting the high performance computing methods. In my project, I am trying to distribute process and getting the optimize result to Job processing can distributed and assign the task which one is memory intensity or CPU intensity. There are some gaps in the previous research that I have found in my research. The fruitful advantage of this method is that larger number of member node might be alive for longer time and Job profiling can assign the task.

- **FLOWCHART OF PROPOSED WORK:**



PROPOSED ALGORITHM:

A Client side request:
(Job Processing Request)

```
Step1: Request from client for job processing
Step 2: Job profiling for identify
            CPU/ Memory intensity jobs
Step 3: Assign Priority to data nodes
           High-CPU & High-Memory
Step 4: If job is CPU intensity
            Assign jobs => High_CPU nodes
     else if  job is memory intensity
                  Assign jobs => High_Memory nodes
Step 5: Collect results.
```

V. SIMULATION AND RESULTS

In this project we will doing a set-up of hadoop node cluster. As per mention here in hadoop have a two or more include a "DataNodes in a distributed Hadoop environment" here we are bring two machines to be used such as-Master & Slave, On both of these machine can be run on the datanode.

A. Prerequisites:

• A platform can develop and produce to be supported "GNU/Linux".

• "Java™ must be installed"-theses is recommended to "Java versions are described at HadoopJavaVersions"

○ "ssh" must be installed&sshd must be "running to use the Hadoop scripts" to be manage a remote "Hadoop-daemons"

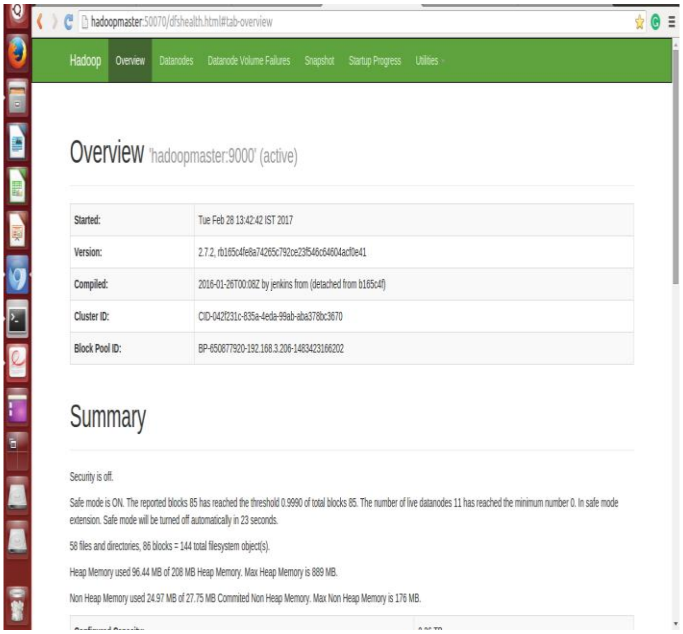
▪ "Cent OS-6.5"

▪ "Hadoop-2.7.3"

▪ "JAVA 8"

• "SSH"

Simulation Results:



Overview 'hadoopmaster:9000' (active)

Started:	Tue Feb 28 13:42:42 IST 2017
Version:	2.7.2, rb165c4fe8a74265c792ce23f546c64604acf0e41
Compiled:	2016-01-26T00:08Z by jenkins from (detached from b165c4f)
Cluster ID:	CID-042f231c-835a-4eda-99ab-aba378bc3670
Block Pool ID:	BP-650877920-192.168.3.206-1483423166202

Summary

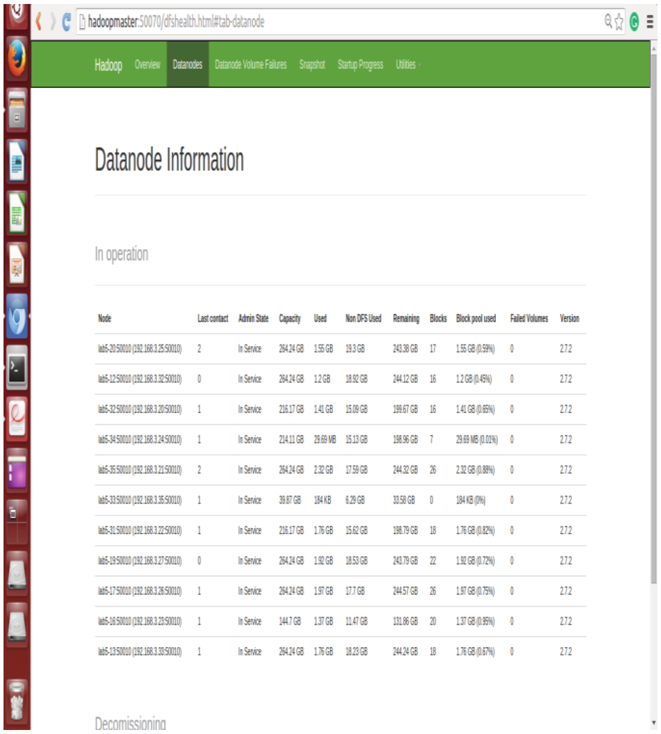
Security is off.

Safe mode is ON. The reported blocks 85 has reached the threshold 0.9990 of total blocks 85. The number of live datanodes 11 has reached the minimum number 0. In safe mode extension. Safe mode will be turned off automatically in 23 seconds.

58 files and directories, 86 blocks = 144 total filesystem object(s).

Heap Memory used 96.44 MB of 208 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 24.97 MB of 27.75 MB Commited Non Heap Memory. Max Non Heap Memory is 176 MB.



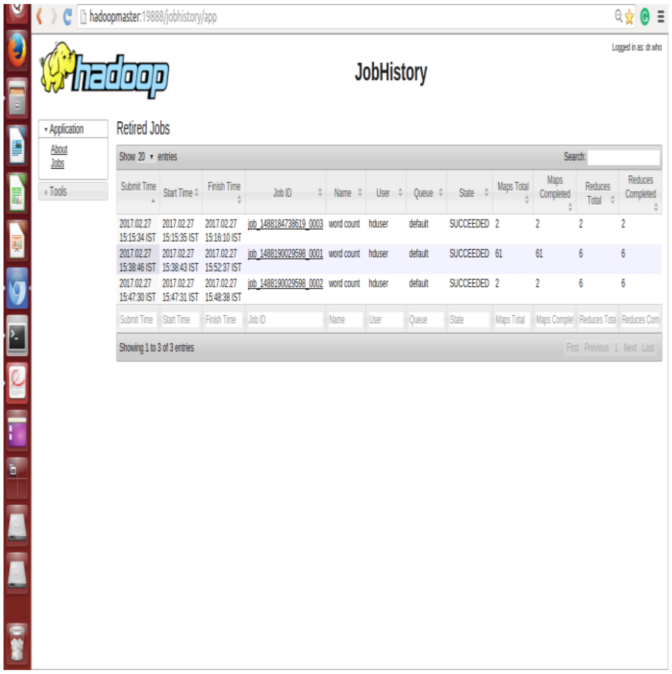
hadoopmaster:50070/dfshealth.html#tab-datanode

Datanode Information

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
lab5-20:50010 (192.168.3.25:50010)	2	In Service	264.24 GB	1.55 GB	19.3 GB	243.38 GB	17	1.55 GB (0.59%)	0	2.7.2
lab5-12:50010 (192.168.3.32:50010)	0	In Service	264.24 GB	1.2 GB	18.92 GB	244.12 GB	16	1.2 GB (0.45%)	0	2.7.2
lab5-32:50010 (192.168.3.20:50010)	1	In Service	216.17 GB	1.41 GB	15.09 GB	199.67 GB	16	1.41 GB (0.65%)	0	2.7.2
lab5-34:50010 (192.168.3.24:50010)	1	In Service	214.11 GB	29.69 MB	15.13 GB	198.96 GB	7	29.69 MB (0.01%)	0	2.7.2
lab5-35:50010 (192.168.3.21:50010)	2	In Service	264.24 GB	2.32 GB	17.59 GB	244.32 GB	26	2.32 GB (0.88%)	0	2.7.2
lab5-33:50010 (192.168.3.35:50010)	1	In Service	39.87 GB	184 KB	6.29 GB	33.58 GB	0	184 KB (0%)	0	2.7.2
lab5-31:50010 (192.168.3.22:50010)	1	In Service	216.17 GB	1.76 GB	15.62 GB	198.79 GB	18	1.76 GB (0.82%)	0	2.7.2
lab5-19:50010 (192.168.3.27:50010)	0	In Service	264.24 GB	1.92 GB	18.53 GB	243.79 GB	22	1.92 GB (0.72%)	0	2.7.2
lab5-17:50010 (192.168.3.26:50010)	1	In Service	264.24 GB	1.97 GB	17.7 GB	244.57 GB	26	1.97 GB (0.75%)	0	2.7.2
lab5-16:50010 (192.168.3.23:50010)	1	In Service	144.7 GB	1.37 GB	11.47 GB	131.86 GB	20	1.37 GB (0.95%)	0	2.7.2
lab5-13:50010 (192.168.3.33:50010)	1	In Service	264.24 GB	1.76 GB	18.23 GB	244.24 GB	18	1.76 GB (0.67%)	0	2.7.2

Decomissioning



hadoopmaster:19888/jobhistory/app

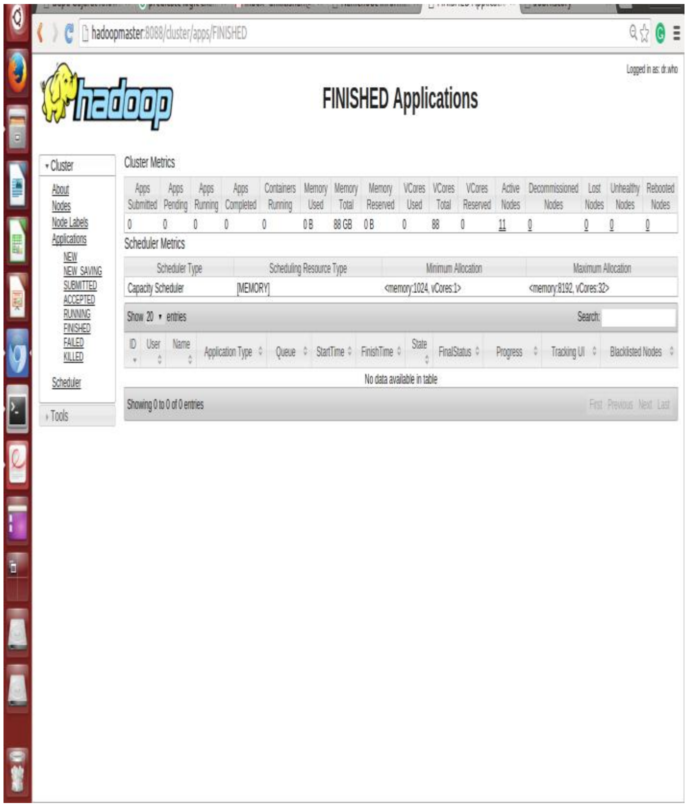
JobHistory

Retired Jobs

Show 20 entries Search:

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed
2017.02.27 15:15:34 IST	2017.02.27 15:15:35 IST	2017.02.27 15:16:10 IST	job_1488184738619_0003	word count	hduser	default	SUCCEEDED	2	2	2	2
2017.02.27 15:38:46 IST	2017.02.27 15:38:43 IST	2017.02.27 15:52:37 IST	job_1488190029598_0001	word count	hduser	default	SUCCEEDED	61	61	6	6
2017.02.27 15:47:30 IST	2017.02.27 15:47:31 IST	2017.02.27 15:48:38 IST	job_1488190029598_0002	word count	hduser	default	SUCCEEDED	2	2	6	6

Showing 1 to 3 of 3 entries



hadoopmaster:8088/cluster/apps/FINISHED

FINISHED Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	88 GB	0 B	0	88	0	11	0	0	0	0

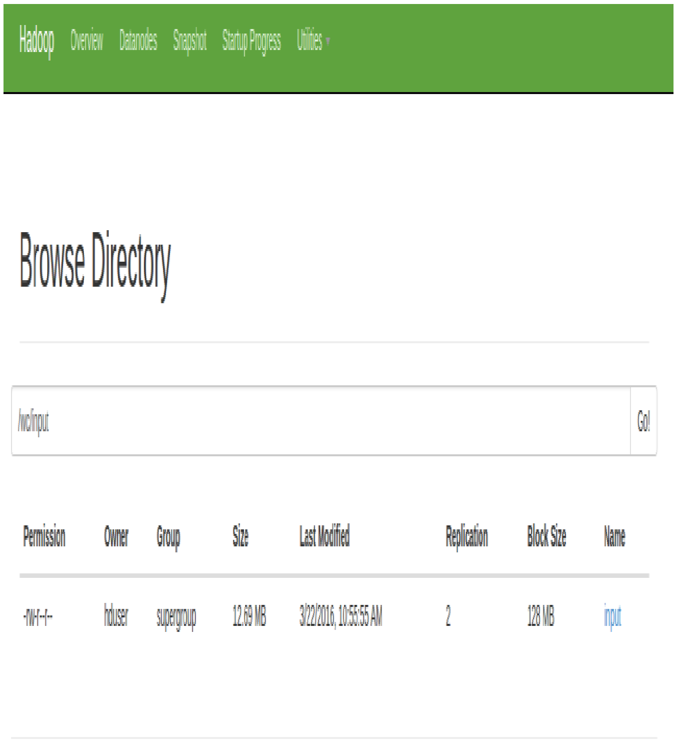
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:32>

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
No data available in table											

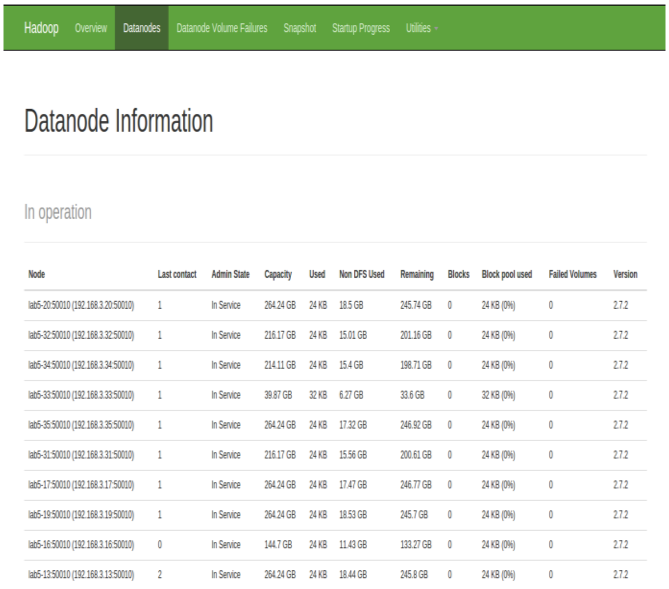
Showing 0 to 0 of 0 entries



Browse Directory

/wcinput Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hduser	supergroup	12.69 MB	3/22/2016, 10:55:55 AM	2	128 MB	input



Datanode Information

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
lab5-20:50010 (192.168.3.20:50010)	1	In Service	264.24 GB	24 KB	18.5 GB	245.74 GB	0	24 KB (0%)	0	2.7.2
lab5-32:50010 (192.168.3.32:50010)	1	In Service	216.17 GB	24 KB	15.01 GB	201.16 GB	0	24 KB (0%)	0	2.7.2
lab5-34:50010 (192.168.3.34:50010)	1	In Service	214.11 GB	24 KB	15.4 GB	198.71 GB	0	24 KB (0%)	0	2.7.2
lab5-33:50010 (192.168.3.33:50010)	1	In Service	39.87 GB	32 KB	6.27 GB	33.6 GB	0	32 KB (0%)	0	2.7.2
lab5-35:50010 (192.168.3.35:50010)	1	In Service	264.24 GB	24 KB	17.32 GB	246.92 GB	0	24 KB (0%)	0	2.7.2
lab5-31:50010 (192.168.3.31:50010)	1	In Service	216.17 GB	24 KB	15.56 GB	200.61 GB	0	24 KB (0%)	0	2.7.2
lab5-17:50010 (192.168.3.17:50010)	1	In Service	264.24 GB	24 KB	17.47 GB	246.77 GB	0	24 KB (0%)	0	2.7.2
lab5-19:50010 (192.168.3.19:50010)	1	In Service	264.24 GB	24 KB	18.53 GB	245.7 GB	0	24 KB (0%)	0	2.7.2
lab5-16:50010 (192.168.3.16:50010)	0	In Service	144.7 GB	24 KB	11.43 GB	133.27 GB	0	24 KB (0%)	0	2.7.2
lab5-13:50010 (192.168.3.13:50010)	2	In Service	264.24 GB	24 KB	18.44 GB	245.8 GB	0	24 KB (0%)	0	2.7.2

VI. CONCLUSION

- Hadoop has been seen as a technology which has been evolved and improved over the years. Hadoop provides an environment for distributed computing. Job scheduling is seen as an important aspect of high performance in Hadoop cluster. In our experiment, request from client side for job processing. A job Profiling can be identified in which one is high CPU intensity or Memory intensity. A Job Profiling task is distributing the process or an application as a High CPU or Memory intensity and assign the task as a priority based data node.

- If the job is high CPU intensity, then job processing should assign a job in High CPU nodes, else the job is Memory intensity then job processing assigns in Memory nodes. After all nodes assigned as a priority based then collect result which nodes can perform in high CPU/Memory node.

REFERENCES

[1]. Ahmed Abdulhakim Al-Absi, Dae-Ki Kang and Myong-Jong Kim, “Enhancing Dataset Processing in Hadoop YARN Performance for Big Data Applications” Springer, 2016.

[2]. Wei Dai, Ibrahim Ibrahim, Mostafa Bassiouni, “A New Replica Placement Policy for Hadoop Distributed File System” IEEE , 2016.

[3]. Ankita Saldhi, Abhinav Goel, Dipesh Yadav, Ankur Saldhi, Dhruv Saksena, S. Indu, “Big Data Analysis Using Hadoop Cluster” IEEE, 2014.

[4]. Md. Wasi-ur-Rahman, Nusrat Sharmin Islam, Xiaoyi Lu, Jithin Jose, Hari Subramoni, Hao Wang and Dhabaleswar K. (DK) Panda, “High-Performance RDMA-based Design of Hadoop MapReduce over InfiniBand” IEEE, 2013.

[5]. Zujie Ren, Xianghua Xu, Jian Wan, “Workload Characterization on a Production Hadoop Cluster: A Case Study on Taobao” IEEE,2012.

[6]. Xiao Yang, Jianling Sun, “AN ANALYTICAL PERFORMANCE MODEL OF MAPREDUCE” IEEE, 2011.

[7]. Mukesh Singla, “A Brief Preview of Efficient Hadoop Job Schedulers” Research India Publications, 2016.

[8]. Mukesh Singla, “A survey on Static and Dynamic Hadoop Schedulers” Research India Publications, 2017.

[9]. Yadav Pratik A., Sase Yuvraj S, “Big Data Implementation Using hadoop and grid Computing” IJIRSET-2014.

[10]. Lawal M. A., “Implementing big data management on Grid-Computing environment.”, IJECS-2014.

[11]. Garlasu Dan;Sandulescu Virginia; Mariana.M., “A Big Data implementation based on Grid Computing.”, IJECS-2014.

[12]. Savant Ila;Muke Richa; Narlawar Nilay, “A Study on hadoop_MapReduce_Techniques and Applications on Grid”, IJERT-2013.

[13]. Dr.B.venkata R.Reddy;G.Narendra; R.Navateja R, “Big data accomplishment using apache hadoop & grid computing environment”, IJIRCCE-2016.

[14]. “https://hortonworks.com/blog/hadoop-in-perspective-systems-for-scientific-computing/”

[15]. “https://www.slideshare.net/allenwittenauer/deploying-grid-services-using-apache-hadoop-15342927”

[16]. https://www.sas.com/en_in/insights/big-data/hadoop.html

[17]. http://bigdatank.blogspot.com/2015/05/hadoop-introduction-and-motivation.html

[18]. https://www.ebayinc.com/stories/blogs/tech/secure-communication-in-hadoop-without-hurting-performance/

[19]. https://www.edureka.co/blog/setting-up-a-multi-node-cluster-in-hadoop-2.X

[20]. https://www.tutorialspoint.com/hadoop/hadoop_multi_node_cluster.htm

[21]. https://www.3pillarglobal.com/insights/a-quick-set-up-guide-for-single-node-hadoop-clusters

[22]. https://acadgild.com/blog/hadoop-multinode-cluster-configuration

[23]. https://clusteringformeremortals.com/2009/09/15/step-by-step-configuring-a-2-node-multi-site-cluster-on-windows-server-2008-r2-%E2%80%93-part-1/

[24]. https://data-flair.training/blogs/data-locality-in-hadoop-mapreduce/

[25]. https://intellipaat.com/tutorial/hadoop-tutorial/introduction-hadoop/

AUTHORS PROFILE

Maitry Joshi completed Bachelor of Information Technology from Hasmukh Goswami College of Engineering, Vahelal, Ahmedabad, India in 2015 and Master of Computer Engineering from Shankarsinh Vaghela Bapu Group of Institute, Vasan, Gujarat, India Pursuing.

Prof. Ankit Shah works as an Assistant Professor in Department of Computer Engineering and Technology, Shankarsinh Vaghela Bapu Group of Institute and also the Head of the Department. He had completed his P.hd. His main research work focuses on Hadoop.