Parallel programming using Hadoop

CSE 4001 Parallel and Distributed computing

Slot D1

Reg.No.	Name
15BCE0290	VARSHINI S
15BCE0491	DEVI PRATYUSHA
15BCE0685	JANDHYALA KATYAYANI

Introduction

Parallel computing deals with the software and the system architectural issues referred to the concurrent execution of applications. Parallel programming has become an area of highly active research interest over decades. It is now always focusing on the high performance computing.

Parallel computing has become region of interest for the electrical and the electronics field of sciences. The reason behind this is solely being the shift of semiconductor industry to the multi-core processors.

With the multiple processors, working side-by-side on shared memory, the shared memory multiprocessors came into the picture which dates back to early 1960's or late 1950's. Then came the massive parallel processors (MPPs) ,in the mid 1980's , they were showing the outstanding performance . Ever since this, the growth of MPPs continued to enhance at a consistent pace over size and power.

Further in the late 80's, clusters emerged as a competition displacing the MPPs to the entirely greater extent for numerous applications. Clusters are recognized as a category of parallel computer built from a number of off-the-shelf computers connected by and with the off-the-shelf network. In this modern era of information, clusters are the job doers of the scientific computations and most prevalent framework in the data centers.

Literature Survey

MapReduce is a parallel programming model and a related usage presented by Google. Client discusses about the two factors specifically. The Map and the Reduce.

The underlying MapReduce library parallelizes, automatically, the calculations and handles entangled issues like information conveyance, stack adjusting and adaptation to internal failure(fault tolerance)in [4]. Huge input datasets that spread across several machines, need to be parallelized. Hadoop is an open-source counterpart of the original MapReduce implementation by Google. Parallelizing computing in large clusters is the center of interest of Hadoop and MapReduce.

This paper[7] is an attempt to illustrate the MapReduce programming model and its applications. The main aim of this program is to portrait the work process of MapReduce. Some essential issues, similar to adaptation to non-critical failure, are contemplated in more detail. Indeed, even the representation of working of Map Reduce is given.

Given an information sensitive application [7],[10] running on a Hadoop Map Reduce cluster, the approach of this paper attempts to explain how information position is done in Hadoop Architecture.

Overview of current scenario

At present scenario, the Apache Hadoop project is the most prevalent implementation of MapReduce. Also, it manages all the appropriate details required to scale the MapReduce operations. With the business support and group commitments over years Hadoop is now recognized as the completely-featured, extensible information processing-data handling platform. There are scores of other open source ventures composed particularly to work with Hadoop. Apache Pig and Cascading, for example, give abnormal state dialects and deliberations for information control. Apache Hive gives an information distribution center over Hadoop.

As the Hadoop system abandoned the opposition, organizations like Microsoft, who were attempting to fabricate their own particular MapReduce stage, in the end it chose to help Hadoop under the insistence of client request. Apart from these,

powerhouses like Facebook, Netflix, LinkedIn, and Yahoo have been utilizing Hadoop for a considerable latency and time span. TRUECar, is another Hadoop client in the business, which has cost of quarter-dollar per GB with Hadoop, which is quite not a case when they weren't using Hadoop.

In a way, virtualizing Hadoop is the subject of some debate among Hadoop architects. The execution of virtualized Hadoop and its performance are most talked about in the present day scenario.

Hadoop Distributed File System

The HDFS is a scalable, distributed and portable file system written in Java for the Hadoop frameworkthe Hadoop structure. Some consider it to rather be an information store because of its absence of POSIX consistence, yet it provides shell orders and Java application programming interface (Java API) strategies that are like other storage frameworks. A Hadoop cluster has ostensibly a solitary namenode in addition to a group of datanodes, in spite of the fact that excess choices are accessible for the namenode because of its criticality. Each datanode serves up squares of information over the system utilizing a piece convention particular to HDFS. The document framework utilizes TCP/IProtocol attachments for all correspondence. As well customers utilize remote strategy calls (RPC) to interact.

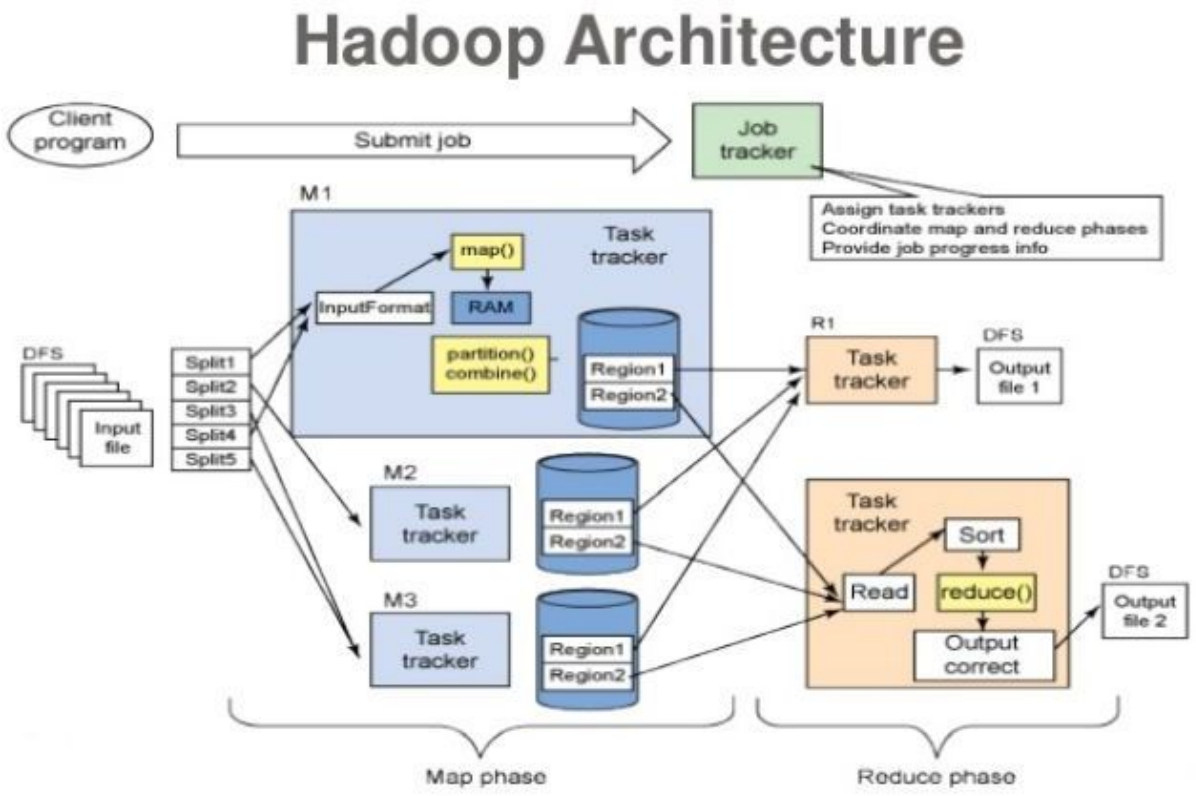

Fig 1 depicts the entire discussion

Conclusion

In this implementation, we try to explain the working of parallel processing. It also includes illustrating the working of Map Reduce framework in the Hadoop Distributed File System(HDFS). The prime idea is that the Map Reduce framework modifies and reduces the complexity of the running distributed data processing operations parallelly, across multiple nodes in a cluster.

Any developer with no particular information of conveyed parallel programming will be able to create a MapReducer.

The adaptation to internal failure highlight is actualized by the Map Reduce by utilizing Replication. Hadoop accomplishes adaptation to fault tolerance by methods

for data replication. The main aim of this project is to study, understand and implement parallel processing for clusters using MapReduce

References

[1] http://developer.yahoo.com/hadoop/tutorial/module3.html

[2] Shweta Pandey, Vrinda Tokekar. The prominence of MapReduce in BIG DATA Processing. In Fourth International Conference on Communication Systems and Network Technologies,IEEE, pages 555-560 , 2014

[3] Google's MapReduce Programming Model-Revisited_Ralf L¨ammel

[4] http://ieeexplore.ieee.org: Improving MapReduce performance through data placement in
heterogeneous Hadoop clusters

[5] https://www.tutorialspoint.com/map_reduce/implementation_in_hadoop.htm

[6]Simplified data processing on large clusters.
Commun. ACM, 51(1):107–113, 2008. Jeffrey Dean and Sanjay Ghemawat. Mapreduce

[7] Review of Distributed File Systems: Concepts and Case Studies ECE 677 Distributed Computing
Systems

[8] Levy E. and Silberschatz A., "Distributed FileSystems: Concepts and Examples"

[9] Yahoo Research-http://dimacs.rutgers.edu/Workshops/Parallel/slides/suri.pdf

[10]
http://www.semantikoz.com/blog/hadoop-index-optimise-map-reduce-data-access-resultin-
32x-speedup/

[11]
http://www.drdobbs.com/parallel/indexing-and-searching-on-a-hadoop-distr/226300241

[12] Source from AIAA 2011: Survey of Parallel Data Processing in Context with MapReduce by
Madhavi Vaidya