ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 4, Issue 10, October 2017

Hybrid Cloud Environment using Map reduce for Effective Job Scheduling

[1] G.Suhasini, [2] Dr. P.Niranjan
[1] Research scholar, [2] Rresearch supervisor
[1][2] mewar university, Chittorgarh, Rajasthan

Abstract: **MapReduce plays major role in data intensive applications. Hadoop MapReduce is wide used framework across the world also used for data-intensive applications that require exploiting data processing power of distributed programming frameworks. One of the great options of Hadoop MapReduce is its support for cloud organizations. Many service providers like Amazon with Elastic MapReduce are having provision for running Hadoop applications in this context, it's essential to have effective job planning and resource provisioning mechanisms for public cloud. Khan et al. proposed Hadoop Performance Modelling for job estimation and resource provisioning to improve user satisfaction besides helping users to have optimal utilization of cloud resources. This paper proposed effective job scheduling and resource provisioning The users are made responsible to make decisions on resource needs. Often the users are unaware of the resource requirements. This proposed methodology has provision for estimating the execution time of given job which has deadline requirements. In a hybrid cloud environment, the estimation of time is made. This can help in automating resource provision and job scheduling to ensure that job is execution in the given deadline.**

Index Terms— Cloud computing, Hadoop MapReduce, performance modeling, job estimation, resource provisioning

INTRODUCTION

Nowadays Many Organizations in the world wide are capturing and analyzing data continuously. This datasets collects from various Social networking, World Wide Web (WWW) and wireless networks are producing data continuously. This unstructured data analyzing is big task in cloud environment .When data is accumulated in huge quantities and growing in size, it assumes attributes of big data. Analyzing such data is very important for organizations to make well informed decisions. Based on this analyzing data criteria organizations runs their business .if not analyzing data with any reason organization losses business. To process this hug amount of data Google introduced new programming model[7] name MapReduce. This MapReduce plays an important role in data sensitive applications.. The rationale behind this is that the MapReduce [10] framework is highly scalable, data parallel model and fault-tolerant in nature. There are many Map Reduce implementations such as Hadoop, Dryad, Phoenix, and Mars among them Hadoop is widely used framework across the globe. These are used for data-intensive applications that need to exploit parallel processing power of distributed programming frameworks[3]. Hadoop Map Reduce is its support for cloud computing. Lot off small and medium organizations are interacting with clouds to reduce their cost investments. So daily number of datasets collecting from various organizations. Present different providers are available like Amazon with EMR to run Hadoop applications .Recently Khan et al proposed hadoop performance modeling for job estimation and resource provisioning. This cloud can improve user satisfaction besides helping users to have optimal utilization of cloud resources. Main drawback is the over-provisioning of resourcing for user jobs with large deadlines in the cases where VMs are configured with a large number of map slots and reduce slots.

RELATED WORKS

Hadoop execution demonstrating is a developing theme that arrangements with work streamlining, planning, estimation and asset provisioning. As of late this point has gotten an incredible consideration from the exploration group and various models have been proposed. Morton et al. proposed the parallax show and later the Para Timer display that gauges the execution of the Pig parallel questions, which can be converted into understanding of MapReduce occupations. They utilize troubleshoot keeps running of a similar inquiry on input information tests to anticipate the relative advance of the guide and lessen stages. This work depends on streamlined suppositions that the terms of the guide undertakings and the diminish errands are the same for a MapReduce application. In any case, in actuality, the lengths of the guide undertakings and the lessen errands can't be the same in light of the fact that the spans of these assignments are relied upon various variables. All the more critically, the spans of the decrease errands in covering and non-covering stages are altogether different. Ganapathi et al. [9] utilized a multivariate Kernel Canonical Correlation Analysis relapse procedure to foresee the execution of Hive inquiry. Be that as it may, their aim was to demonstrate the material ness of strategy with regards to MapReduce. Kadirvel et al. [10] proposed Machine Learning strategies to foresee the execution of Hadoop employments. In any case, this work does not have an exhaustive numerical model for work estimation. Lin et al. [1] proposed a cost vector which contains the cost of plate I/O, organize activity, computational multifaceted nature, CPU and inner sort. The cost vector is utilized to evaluate the execution lengths of the guide and diminish errands. It is trying to precisely gauge the cost of these elements in a circumstance where different assignments go after assets. Besides, this work is just assessed to evaluate the execution times of the guide errands and no estimations on decrease assignments are introduced. The later work [2] considers asset conflict and undertakings disappointment circumstances. A test system is utilized to assess the viability of the model. In any case, test system base methodologies are conceivably mistake inclined in light of the fact that it is trying to plan a precise test system that can exhaustively mimic the interior flow of complex MapReduce applications. Virajith et al. [3] proposed a framework called Bazaar that predicts Hadoop work execution and arrangements assets in term of VMs to fulfill client necessities. The work displayed in [4] utilizes the Principle Component Analysis procedure to advance Hadoop employmentsin lightof different design parameters. In any case, these models forget both the covering and non-covering phases of the rearrange stage. There is collection of work that spotlights on ideal asset provisioning for Hadoop employments. Tian et al. [5] proposed a cost demonstrate that gauges the execution of work and arrangements the assets for the activity utilizing a basic relapse strategy. Chen et al. [6] additionally enhanced the cost display which utilizes the animal power look strategy for provisioning the ideal group assets in term of guide spaces and decrease openings for Hadoop occupations. The proposed cost show can foresee the execution of an occupation and arrangements the assets required. In any case, in the two models , the quantity of lessen undertakings must be equivalent to the quantity of reduce spaces which implies that these two models just consider a solitary flood of the decrease stage. It is questionable that a Hadoop work performs better when various influxes of the reduce stage are utilized as a part of examination with the utilization of a solitary, particularly in conditions where a little measure of assets is accessible yet handling a vast dataset.

METHODOLOGY

In this Paper we proposed a new effective job scheduling and resource provisioning to overcome khan eal.prolems.In this methodology users can make decisions on resource requirements this will make user satisfaction it's also improves system performance .. In this research, the methodology has provision for estimating the execution time of given job which has deadline requirements. In a hybrid cloud environment, the estimation of time is made. This can help in automating resource provision and job scheduling to ensure that job is execution in the given deadline. Before looking at the proposed methodology, Figure shows the MapReduce work flow. As per figure 1 shows exccution model of MapReduce its starts with a dataset given as input. After given input Map Phase model will strat. The Map phase consistes multiple Map tasks that act on given data in parallel and produce intermediate output. The intermediate dataset is subjected to shuffling or sorting. Then the output of shuffling is given to reduce phase. In the reduce phase, the final output is generated and stored in Hadoop Distributed File System (HDFS).

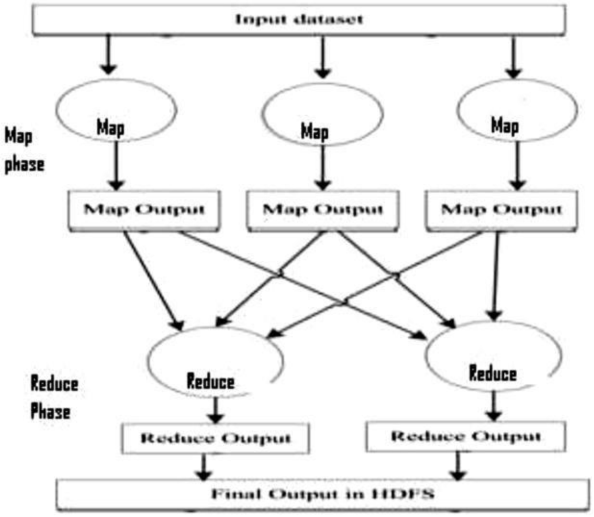

Fig 1:MapReduce Execution Model

CONCLUSION:

This proposed methodology has provision for estimating the execution time of given job which has deadline requirements. In a hybrid cloud environment, the estimation of time is made. This can help in automating resource provision and job scheduling to ensure that job is execution in the given deadline.