STUDY ON JOB SCHEDULING AND RESOURCE PROVISIONING MODELS FOR EFFICIENT MAPREDUCE PROGRAMMING FOR BIG DATA PROCESSING

[1]G Suhasini, [2]Dr.C D Kumawat
Research Scholar Mewar University,Chittorgarh,Rajasthan
Research Supervisor Mewar University,Chittorgarh,Rajasthan

Abstract.
Many companies are increasingly using MapReduce for efficient large scale data processing such as personalized advertising, spam detection, and different data mining tasks. Cloud computing offers an attractive option for businesses to rent a suitable size Hadoop cluster, consume resources as a service, and pay only for resources that were utilized. One of the open questions in such environments is the amount of resources that a user should lease from the service provider. Often, a user targets specific performance goals and the application needs to complete data processing by a certain time deadline. However, currently, the task of estimating required resources to meet application performance goals is solely the users' responsibility. In this work, we introduce a novel framework and technique to address this problem and to offer a new resource sizing and provisioning service in MapReduce environments. For a MapReduce job that needs to be completed within a certain time, the job profile is built from the job past executions or by executing the application on a smaller data set using an automated profiling tool. Then, by applying scaling rules combined with a fast and efficient capacity planning model, we generate a set of resource provisioning options. Moreover, we design a model for estimating the impact of node failures on a job completion time to evaluate worst case scenarios. We validate the accuracy of our models using a set of realistic applications. The predicted completion times of generated resource provisioning options are within 10% of the measured times in our 66-node Hadoop cluster.

1. INTRODUCTION

Organizations in the real world are capturing data continuously. Of late much importance is given to data which resulted in exponential growth of data. Social networking, World Wide Web (WWW) and wireless networks to mention few are producing data continuously. 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. If, for any reason, the data is not analyzed well, the organization loses opportunities in business arena. In order to process huge amount of data, a new programming model by name MapReduce [1] was introduced by Google. Right from the inception, MapReduce became a reliable model for processing big data. The rationale behind this is that the MapReduce framework is highly scalable, data parallel model and fault-tolerant in nature. There are many MapReduce implementations such as Hadoop, Dryad, Phoenix, and Mars, 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.

One of the good features of Hadoop MapReduce is it support for cloud computing. Thus organizations can use it in pay per use fashion. Due to financial constraints small and medium organizations opt for this kind of model. Many service providers like Amazon with Elastic MapReduce (EMR) are having provision for running Hadoop applications. In this context, it is essential to have effective job scheduling and resource provisioning mechanisms for public cloud. Khan et al. proposed Hadoop Performance Modelling for job estimation and resource provisioning. This could improve user satisfaction besides helping users to have optimal utilization of cloud resources. In order to estimate execution time of a job, they employed Locally Weighted Linear Regression (LWLR). In order to satisfy user jobs with deadline requirements, they employed Lagrange Multipliers technique for optimal resource provisioning. The Hadoop performance model proposed by Khan et al. has certain limitations. The 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. In this research we focus on the over-provisioning problem where dynamic overhead of VMs is considered in order to minimize resource over-provisioning. This research also considers multiple Hadoop jobs with deadline requirements for making a model for job scheduling and resource provisioning for efficient programming for big data processing.

2. LITERATURE REVIEW

This section provides review of literature on job scheduling and resource provisioning in cloud computing with respect to MapReduce programming paradigm. Assuncao et al. [1] explored different trends in cloud computing and big data analytics. They specified four aspects of analytics such as architectural support and data management, model development, user interaction and visualization, and business models. They categorized

DOI:10.9756/INTJECSE/V14I6.339 ISSN: 1308-5581 Vol 14, Issue 06 2022

analytics into descriptive, predictive and perspective. Descriptive model is used for modelling past behaviour. Predictive is for forecasting based on the existing data and perspective is for decision making and assessing actions. They also discussed about the characteristics of big data such as variety, velocity, volume, veracity and value. Variety refers to data types. Velocity refers to data production and processing speed. Volume refers to the size of data. Veracity refers to data reliability and trust. Value refers worth obtained from big data after analysis.

Job scheduling plays important role in MapReduce programming. Mashayekhy et al. [2] studied the concept of energy-aware scheduling of MapReduce jobs for big data applications. They proposed and implemented a framework to have energy-aware job scheduling for big data applications with service level agreements (SLA). They also proposed two heuristic algorithms known as energy-aware MapReduce scheduling algorithms. They take care of assignment of map and reduce tasks in order to optimize the energy consumed for MapReduce programming. They used different algorithms for MapReduce programming in Hadoop distributed environment for finding execution time and energy consumption. They tested algorithms with different workloads and found the efficiency of the proposed algorithms in running map and reduce tasks with energy-awareness.

Ruiz-Alvarez et al. [3] considered cost and job execution time for proposing a framework for optimal resource provisioning for cloud MapReduce and hybrid cloud applications. The made a computational model known as Integer Linear Programming (ILP) for making well informed provisioning decisions. In public clouds there are varieties of resources available. It is difficult to allocate resources dynamically. Based on the user preferences, these researchers found a solution that is optimal. Thus optimal resource provisioning is achieved for MapReduce programming. Monte Carlo simulations are made in hybrid cloud for proof of the concept. Hashem et al. [4] reviews the raise of cloud computing and issues involved in cloud computing. The challenges with respect to research in cloud computing found by them include governance, legal, privacy, heterogeneity, data quality, transformation, data integrity, availability, and scalability. Other issues include data security, data analysis, distributed storage, and data staging [4]. Figure 1 shows data sources, content format, data stores, data staging and data processing related to big data.

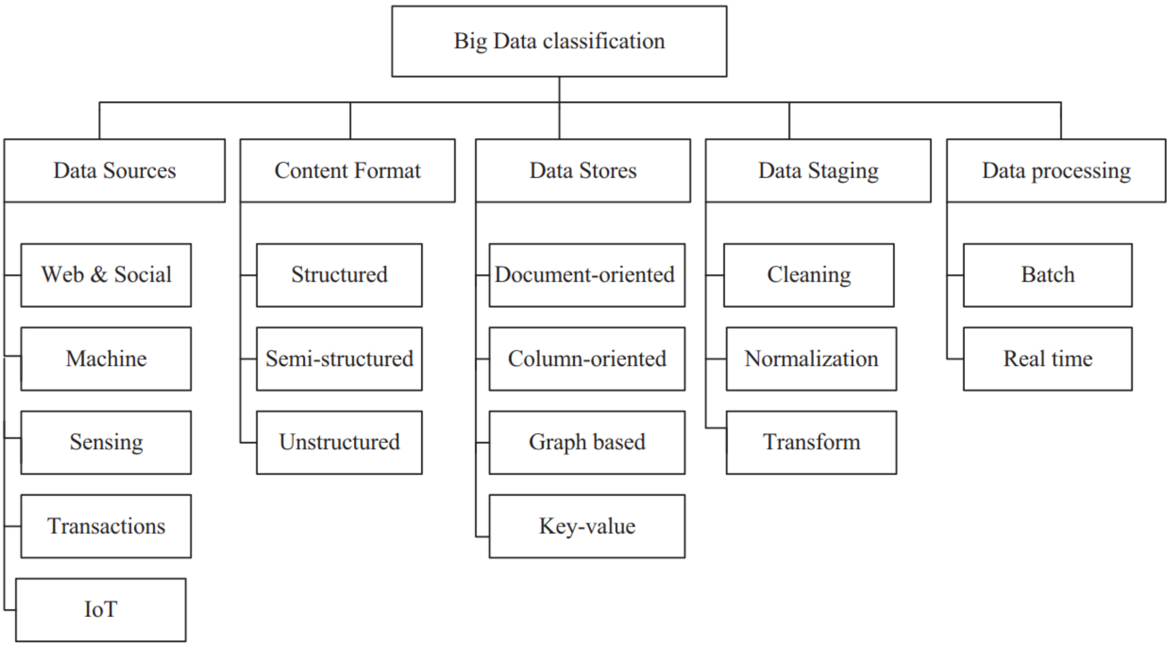


Figure 1: Big data classification

As shown in Figure 4, it is evident that big data comes from different sources such as IoT, transactions, sensing applications, machines, WWW, and social networks. The data has different formats such as structured, unstructured and semi-structured. Data sources of big data include key/value pairs, graph based, column-oriented, and document-oriented. Data staging includes cleaning, normalization and transform. Data processing includes batch processing and real time processing. Kune et al. [5] provided a scheduling solution based on data-awareness and genetic algorithm. They proposed a scheduler based on the genetic algorithm. The data services and computational services are decoupled. The evolutionary approach considers bandwidth, dependencies and computational resources for effective utilization and producing higher throughputs. Their system contains polling of jobs periodically, data discovery, pre-processing, computing resource information, and schedule discovery.

Zhang et al. [6] designed a new model known as CloudFlow programming model for supporting cloud workflow for applications involved in High Performance Computing (HPC). The CloudFlow is data aware and shared data-aware. It supports multiple Map and Reduce functions. Due to the workflow it improves performance of MapReduce programming significantly. It also supports Many Task Computing (MTC) applications with better performance. The job scheduler manages different states of jobs. The states include submission, monitoring and runtime. Bardhan and Menasce [7] presented a model known as Trace Driven Analytic Model (TDAM) for finding the impact of different scheduling algorithms in terms of execution time. Their approach combines both scheduling and queuing in order to have better performance. The components involved in their architecture are job tracker manager, job trace driver, server view manager, task tracker manager, cluster information, residual task calculator, and analytic performance model.

Jha et al. [8] reviewed many data-intensive paradigms such as HPC and Apache Hadoop. They compared different distributed programming frameworks such as Spark, HARP, Python-script, MPI, Mahout, and Hadoop. Zhao et al. [9] focused on workflow for cloud computing. The cloud scientific workflow management is proposed and explored. Their work is compared with different work flow management systems like View, Helper, Vistrails, Taverna, and Swift. Their framework includes task scheduling, scheduling management, cloud resource manager, workflow engines, cloud workflow management service, workflow presentation and visualization, and workflow specification and submission. Cheng et al. [10] explored scheduling in cloud computing known as adaptive task scheduling. This scheduling works in heterogeneous environment for improving MapReduce performance. Their approach is known as Ant which is able to improve job completion time by 23% in Hadoop.

3. RESEARCH METHODOLOGY

The proposed methodology for effective job scheduling and resource provisioning is presented in this section. The MapReduce frameworks and cloud computing paradigms as of now do not provide resource provisioning. The users are made responsible to make decisions on resource needs. Often the users are unaware of the resource requirements. 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 2 shows the MapReduce work flow.

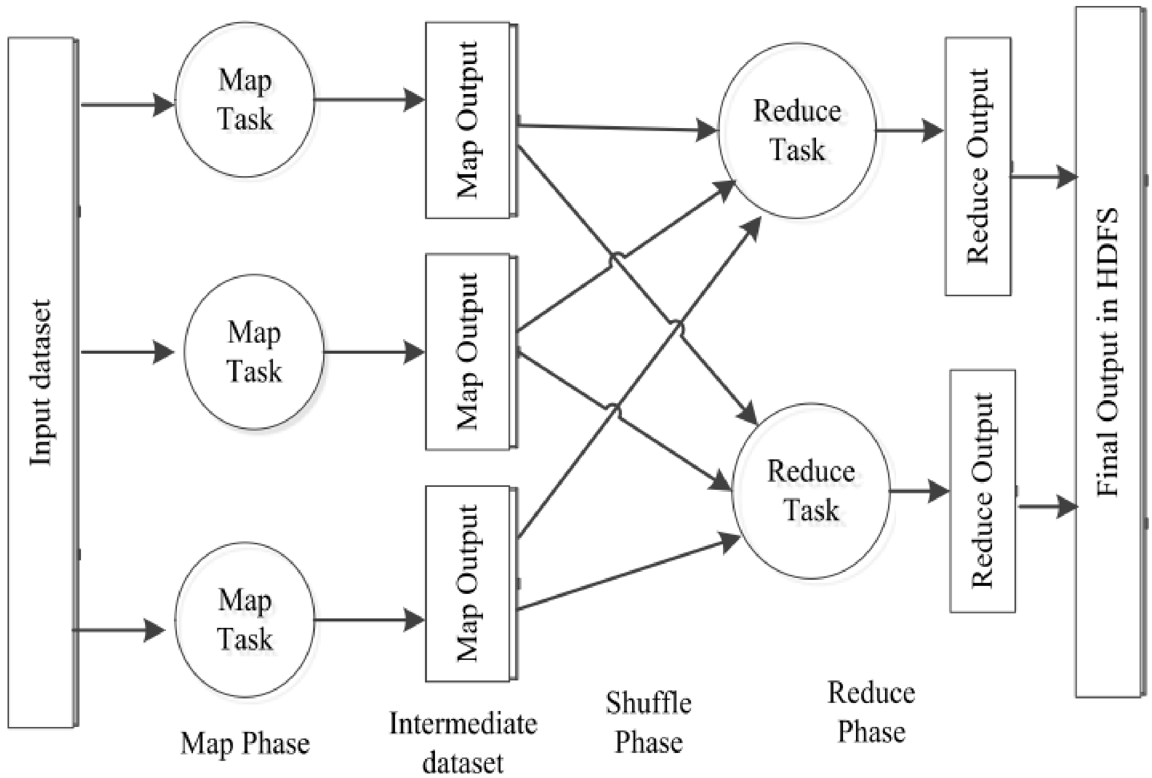
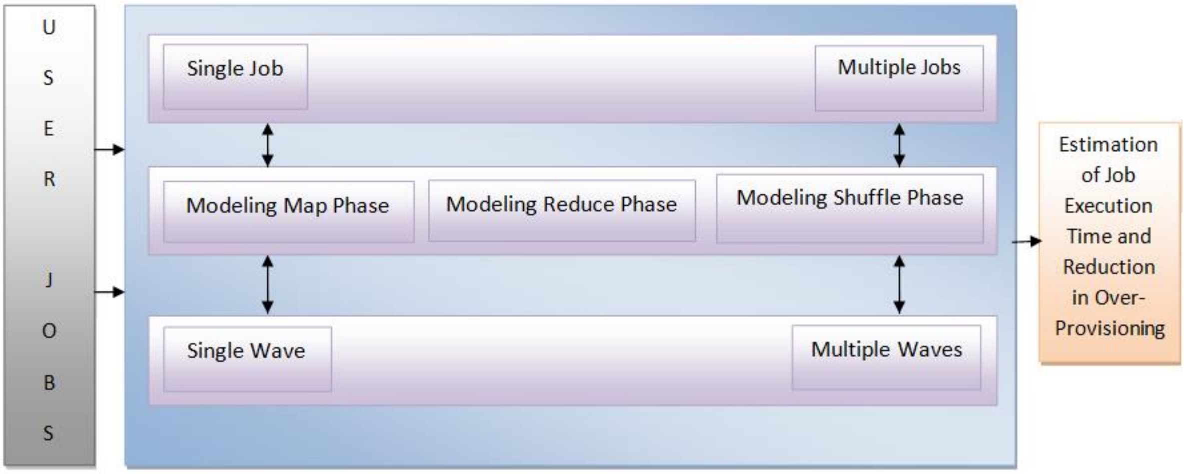


Figure 2: MapReduce execution model

The execution starts with a dataset given as input. Once dataset is given the Map phase starts. The Map phase includes 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).

DOI:10.9756/INTJECSE/V14I6.339 ISSN: 1308-5581 Vol 14, Issue 06 2022



As shown in Figure 3, it is evident that the given jobs are processed with the novel model proposed. The model includes Map phase model, Reduce phase model and Shuffling phase model. In these three phases, the modelling considers single jobs, multiple jobs in the single wave and multiple waves. The aim of the model is to estimate job execution time and have optimized job scheduling and resource provisioning. This model can reduce over-provisioning as it considers the above said environment. It is expected to work fine with jobs with long deadlines with respect to execution.

OBJECTIVES

The objectives identified to fulfil the aim of the proposed research work are as follows.

1. To investigate the present state-of-the-art in job scheduling and resource provisioning in cloud computing to process big data.
2. To propose a novel model for job scheduling and resource provisioning to reduce over provisioning and handle single and multiple users in single wave and multiple wave environments.
3. To implement a prototype application that demonstrates proof of the concept.
4. To evaluate the proposed implementation

SCOPE

The recent innovations on performance modelling of Hadoop or any other distributed programming framework with MapReduce programming paradigm considered estimation of user job estimation time for jobs containing deadline requirements. However, the job scheduling and resource provisioning were made based on the estimation of execution time was not effective for jobs with long deadline requirements. For instance, the Hadoop performance model proposed by Khan et al. has certain limitations. The 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. In this research we focus on the over-provisioning problem where dynamic overhead of VMs is considered in order to minimize resource over-provisioning. This research also considers multiple Hadoop jobs with deadline requirements for making a model for job scheduling and resource provisioning for efficient programming for big data processing.

LIMITATATIONS

The limitations of this research is that it makes use of execution traces of real Hadoop environment. Therefore the evaluation of the work is based on the quality of traces captured from real environment. This research also considers multiple Hadoop jobs with deadline requirements for making a model for job scheduling and resource provisioning for efficient programming for big data processing. Therefore the trace needs to have such details in order to have better evaluation of the work done.

DELIMITATIONS

The proposed research is on the job scheduling and resource provisioning by estimating execution of user jobs with deadline requirements. Since the users do not have knowledge on resource provisioning, the proposed research will have very good impact on the research community, cloud users and service providers. The work is limited to two aspects of MapReduce programming such as job scheduling and resource provisioning.

Conclusion

In this work, we designed a novel framework that aims to enrich private and public clouds offering with an automated SLO-driven resource sizing and provisioning service in MapReduce environments. While there are quite a few companies that offer Hadoop clusters for rent, they do not provide additional performance services to answer a set of typical questions: How much resources the user application needs in order to achieve certain

DOI:10.9756/INTJECSE/V14I6.339 ISSN: 1308-5581 Vol 14, Issue 06 2022

performance goals and complete data processing by a certain time. What is the impact of failures on the job completion time? To answer these questions, we introduced a novel automated profiling technique for MapReduce applications by building a compact but representative job profile in a staging environment. The approach allows executing a given application on the set of the small input datasets. Then by applying a special scaling technique and designed performance models, one can estimate the resources required for processing a targeted large dataset while meeting given SLOs. We also designed a performance model for estimating the impact of failures on MapReduce applications.

REFERENCES

1. D. Assun, N. Calheirosb, Silvia Bianchic, A. S. Nettoc and Rajkumar Buyya. (2014). Big Data Computing and Clouds: Trends and Future Directions. *Journal of Parallel and Distributed Computing*. p1-44.
2. Lena Mashayekhy, MahyarMovahedNejad,DanielGrosu, Quan Zhang and Weisong Shi. (2015). Energy-aware Scheduling of MapReduce Jobs for Big Data Applications. *IEEE*. 20 p1-14.
3. Arkaitz Ruiz-Alvarez, In Kee Kim and Marty Humphrey. (2015). Toward Optimal Resource Provisioning for Cloud MapReduce and Hybrid Cloud Applications. *IEEE*. p1-9.
4. IbrahimAbakerTargioHashem , IbrarYaqoob , NorBadrulAnuar , Salimah Mokhtar , AbdullahGani and SameeUllahKhan. (2015). The riseof “big data” on cloudcomputing:Reviewandopenresearchissues. *InformationSystems*. p1-18.
5. Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige, and Rajkumar Buyya. (2014). Genetic Algorithm based Data-aware Group Scheduling for Big Data Clouds. *IEEE*. p1-9.
6. Fan Zhang, Qutaibah M. Malluhi , Tamer Elsayed , Samee U. Khan, Keqin Li and Albert Y. Zomaya. (2015). CloudFlow: A data-aware programming model for cloud workflow applications on modern HPC systems. *Computer Systems*. p1-13.
7. ShouvikBardhan and Daniel A. Menas. (2012). A Contention-Aware Hybrid Evaluator for Schedulers of Big Data Applications in Computer Clusters. *IEEE*. p1-9.
8. Shantenu Jha, Judy Qiu, Andre Luckow, Pradeep Mantha and Geoffrey C.Fox. (2014). A Tale of Two Data-Intensive Paradigms: Applications, Abstractions, and Architectures. *IEEE*. p1-8.
9. Yong Zhao, Youfu Li ,ShiyongLu,IoanRaicu and Cui Lin. (2013). Devising a Cloud Scientific Workflow Platform for Big Data. *IEEE*. p1-9.
10. Dazhao Cheng, Jia Rao, Yanfei Guo and Xiaobo Zhou. (2014). Improving MapReduce Performance in Heterogeneous Environments with Adaptive Task Tuning. *ACM*. p1-12.