A Survey of Optimization Resource Scheduling Algorithm for Performance Analysis in Big Data Cloud

Zainab Mohanad Issa[1] & Dr. Shaheda Akthar [2]
[1]Research Scholar, Department of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
[2]Registrar(F.A.C), Dr. Abdul Haq Urdu University, Kurnool, Andhra Pradesh, India.

Received: August 30, 2018 Accepted: October 20, 2018

ABSTRACT *The vast usage of big data applications is increase by tremendous changes in big data development. The big data applications is essential for different industries such as healthcare, social media, enterprise applications and etc,. The resource scheduling algorithms play the vital role in optimization of big data performance analysis. Especially the applications on performance analysis in big data is more demand of optimization of resources. The resource scheduler algorithm is a way to optimize the performance analysis of big data applications in cloud. As far as many scheduling algorithms was introduced over the big data applications, it is challenging task of resource allocation and improve the performance. In this work, we put the centre of attention on present state of the art of resource scheduling algorithms in big data cloud. The insights of research findings of this paper provide valuable information to know how the optimal resource scheduling algorithm will improve the performance analysis of big data applications in cloud.*

Keywords: *Big Data, Cloud Computing, Resource Scheduling, Storage*

1. Introduction

The big data clouds is a new emerging trend to perform data performance analysis by using cloud computing as back end technologies for information mining, decision making and knowledge discovery on empirical tools. The big data contains the complex and large amount of data and unable to process with traditional data processing methods. Many researchers are interesting in big data for integration of techniques such as decision making by uncovering hidden patterns , unknown correlations and insights of big data[1]. The big data is greatly inspire the users from different data and application domains to explore data analytics techniques and solutions to get better results in decision making and problem solving. The big data have different type of data such as structured, unstructured and semi-structured data. To manage an analyze these type of data is very difficult with help of traditional database techniques. Big data is dynamic changer for many applications. The term big data means really a big data it large amount of data sets, those are structured and un-structured format. Big data is defined to process large or complex data sets , which is not possible with conventional data process applications. Basically big data consists of data analyzing, data capturing, data creation, data storage, data transfer and privacy information. The big data analysis is a process of examine variety of data in large data sets. The big data is emerging technology it useful for different organizations such as social media, business intelligence and government sectors.

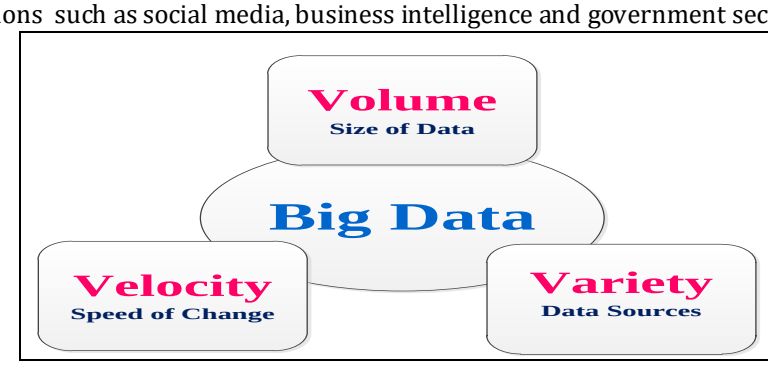

Fig 1. Forms of Big Data

Volume:

Many organizations collects the data from different resources, which consist of different business transactions, social media and information technology.

Velocity:
Data streams unparalleled speed of velocity and have improved in timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in real time operations.

Variety:
Structured data collected from different varieties and numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. We can observe the forms big data in Fig1.

The main importance of big data is reduce the cost and time for analyze the large amount of data. It very helpful in decision making. In big data we have different structured data and different class of technologies. Big data is not a data , it involves various tools , techniques and frameworks[2].

Cloud Computing

Cloud computing is a service provided by the using the internet. The cloud computing is on demand service, it shared resources among different devices. The cloud user can use different facilities such as processing units , memory, servers, operating systems and different environments for application development. In cloud resources can scale upward and downward instantly. With the help of resource virtualization the cloud user can develop , deploy and manage their applications. In the field of information technology the cloud computing is emerging technology[3]. In cloud computing we have three main type of cloud environments available those are 1) private clouds 2) public clouds 3) hybrid clouds.

Private Clouds: In private cloud computing the infrastructure is maintained by single organization. In private clouds requires management resources, maintain the significant level service and maintain the virtual business environment. In private clouds development of every step is security concern and must be addressed and also prevent attacks. The private clouds needs physical attention on control access, hardware management and resource allocation. The periodical refreshment is requires, it leads to additional investment of capital. But the private clouds improve the resource utilization.

Public Clouds: The cloud computing services are provided by the network for public users is called "Public Clouds". The architecture of public cloud is same as private cloud, but it maintained by the third party vendors and provided to users. The public cloud service is widely used for development, deployment and management of applications. The public clouds have high scalability, reliability and security is the main concern[4].

Hybrid Clouds: The hybrid cloud is a combination of private cloud and public cloud. We can be find model diagram of different cloud environments in fig 2.

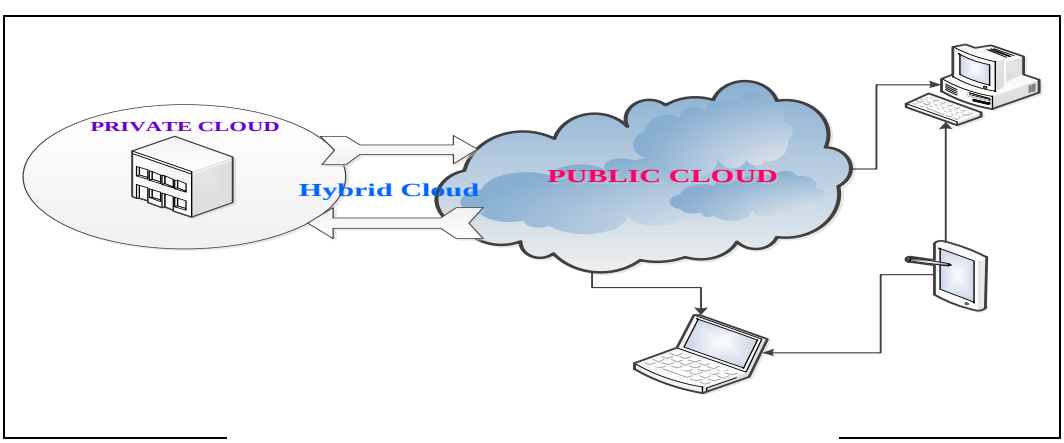

Fig 2. Model diagram of different cloud environments.

In cloud computing the cloud vendors provide their services to cloud users. The cloud users can access their services on pay per use basis. Basically the cloud services are divided into three type of layers.

Platform as a Service(PaaS): The cloud service providers facilitates several environment features. It is useful to cloud users to build their applications. They provide as absolute platform to develop, deploy and maintain the applications according to the requirements.

Software as a Service(SaaS): The cloud service providers provide different kinds of software services. The cloud user can use any type of software service as per requirement of their application. By this service the users are free from installation of software, maintains of software and purchase of software. These kinds risks taken by the service providers.

Infrastructure as a Service(IaaS): In this service cloud user doesn't know the cloud infrastructure. The user can use services of server and other machines. Mostly virtual machines are used, memory allocation is done by dynamically[3].

These services are very helpful to the users , who can't offers these machinery services personally. In cloud computing virtual machines are play a vital role. The cloud contains multiple virtual machines which can be working together. The virtual machines used in optimum way to give benefits to the cloud provider and user. The Fig 3 depicted model shows the working model of big data cloud.

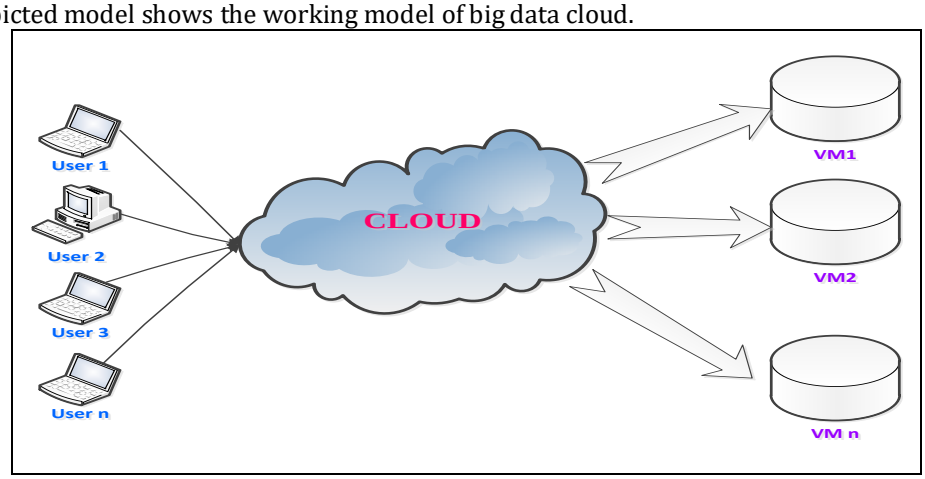

Fig 3. Working Model of Big Data Cloud

The combination of big data and cloud computing services are essential in data analysis. More over the services of these mechanism is increasing vastly. To improve the performance of big data cloud the resource scheduling algorithms are more helpful. In order to make fastest and economical performance, the resource scheduling algorithms are essential.

Table 1. Acronyms and Its Description

S No	Acronym	Description
1	VMs	Virtual Machines
2	BDAA	Big Data Analytic Application
3	ILP	Integer Linear Programming
4	SLA	Service Level Agreement
5	AaaS	Application as a Service
6	QoS	Quality of Service
7	PO	Profit Optimization
8	POS	Profit Optimization Scheduling
9	rCs	Resource Configurations
10	aQs	Admitted Queries
11	HR	Heuristic Resource
12	MDT	Maximum Delay Time
13	aRs	Available Resources
14	eRs	Executable Resources
15	Rs	Existing Resources
16	HRC	Heuristic Resource Configuration
17	lRs	Local-Optimized Resources
18	CPU	Central Processing Unit
19	MI	Processing Requirement of Task
20	MIPS	Processing Capability of Resource

In this paper we reviewed many resource scheduling algorithms that contributed towards improving performance of big data cloud and resource utilization. The remainder of the paper is structured into various sections that throw light on different scheduling algorithms for resource allocation and task scheduling in cloud based big data. It ends with conclusions and recommendations for future work.

2. Profit Optimization Resource Scheduling Algorithm

Yali Zhao et al. (2016) [5] proposed profit optimization resource scheduling algorithm. In this algorithm they formulate resource scheduling algorithm , it overcome scheduling problems and optimize usage of resources. To reach SLA of cloud user the profit optimization resource scheduling algorithm is very useful. The profit optimization resource scheduling algorithm schedule heterogeneous big data cloud resources efficiently and dynamically. This algorithm improve the performance and greatly reduce computation cost in big data cloud. The profit optimization resource scheduling algorithm implemented as follows. The profit optimization resource scheduling algorithm is implemented based ILP formulation. The main aim of the profit optimization algorithm is to maximize resource utilization of big data AaaS in cloud computing. In implementation of this algorithm they consider big data analytic applications and it included different procedures.

The queries from BDAA are send to the profit optimization algorithm. In this first they checks weather any resource configurations are available as per QoS requirements of given query. If rCs available then PO select minimal resource configurations to execute the queries. Then after PO used HR method to pre-decide of optimal heuristic resources as inputs after admitted queries. Here they greatly reduce the space and time. The profit optimization algorithm is greatly utilize the profit optimization scheduling method and feasible solution is applied for schedule resources and queries. If not they utilize maximum delay time method to schedule aQs to HRs to omit the violation of deadline. POs terminates the VMs and scale down resources by periodical chunks to reduce the cost.

HR Procedure

In HR procedure first select available resources from existing resources and execute at least one query. Then it utilizes MDT to map aQs to eRs. MDT is a maximum delay time between the execution time and query deadline. After mapping the queries in aQs is not executed due to limited capacity of eRs. In HR procedure use HRC method to approximate the optimized resources and execute the remaining queries. At finally eRs and IRs combines in HR and make input as POS.

MDT Procedure

In MDT procedure first sort the queries in ascending order and maps query with minimum mdt first to allocate the resource. If any queries are not allocated resources after completion of MDT procedure , The HR utilizes HRC method to estimate the new resources to execute that remaining queries.

HRC Procedure

In this procedure first they applied HS method to find minimized resource configuration. The procedure is executed in following way.

BDAA is requested queries for its set of resource configurations C =(N, E)

N - Represent set of nodes

E - Represents set of edges

Configuration Modification CM = (C_i, C_{i+1})

C_i - Represents resource configuration

The number of resources type available in cloud system is determine by the number CMs available. It is continuous process and execute until find the cheapest resource configuration IRs. Finally POS find optimal resource configuration from HRC generated input resources and greatly reduce memory.

POS Procedure

In this procedure first updates list such as HrS,aQs and BDAA. Then creates the ILP sover to define the objective and define constraints. The constraints are resource capacity, query deadline, query budget and query scheduling time. POS uses the ILP solver to give profit optimization resource scheduling solutions [5].

The profit optimization resource scheduling algorithm achieved better results and greatly reduce computation cost.

3. Task Scheduling Algorithm

The task scheduling algorithm is designed to Saed Abed et al. (2018) [6]. Arrange the resources sharing on all virtual machines. By implementation this algorithm they got high throughput and save the execution time. The task scheduling algorithm followed as, it have multi-layer control nodes. Each control node is responsible for schedule multiple number of compute nodes and this nodes are have fixed number of VMs. The control nodes turn on the data process in optimized way. As per this algorithm the VMs are busy when they are moved from one compute node to another or performing the scheduled task. We can observe the control flow of task scheduling algorithm. In this algorithm they proposed some set of rules , these rules are maintained in each compute nodes. So VMs on each compute nodes are follow the same set of characteristics.

By implementation part this algorithm they configure the characteristics log file at compute nodes instead of VMs. The VMs had dynamic nature can migrate from one compute node to another. On the design of this algorithm the VMs followed resource sharing and perform scheduled tasks as per the log file on the compute nodes[6]. In this way they utilized scheduling resources and performed task with high performance

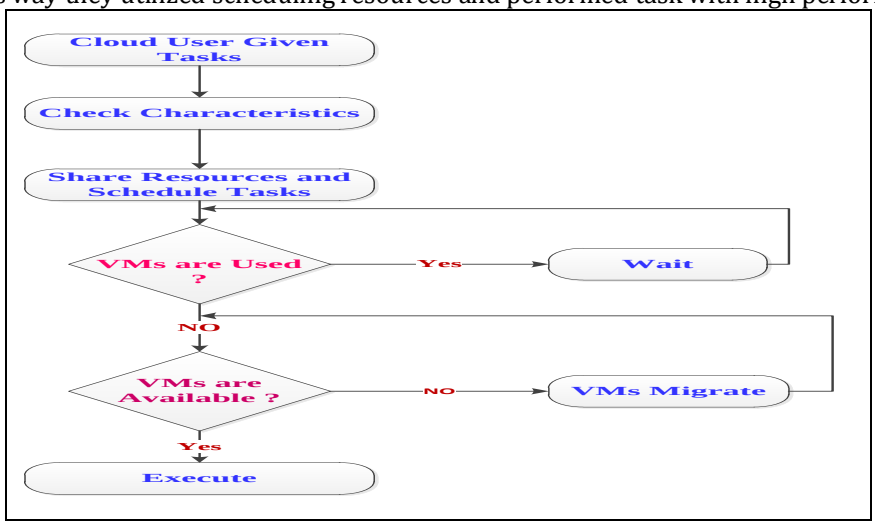

Fig 4. Flow Chart of Task Scheduling Algorithm

4. Smarter Round Robin Scheduling Algorithm

Hicham Gibet Tani et al.(2017) [7] were designed Smarter Round Robin Scheduling Algorithm for resource scheduling and distribution of tasks in cloud based big data. Before going to implementation of this algorithm they discussed about Round Robin Scheduling Algorithm. Actually the round robin scheduling algorithm is simplest and used to share the CPU time among all the scheduled task. In Round Robin Scheduling Algorithm time quantum concept is used for execution of scheduled tasks. But the several implementation of Round Robin based algorithm used static time quantum and dynamic quantum. The static quantum is useful to segments the CPU time for execution of schedule tasks. The static time quantum is not always suitable for all scheduled tasks. The dynamic time quantum is adapt CPU time slices based on changes of scheduled tasks.

For better resource scheduling Smarter Round Robin Scheduling Algorithm is implemented based on dynamic time quantum. The implementation of algorithm is begins with burst time of submitted tasks and queue of waiting list[7].

They followed layered concept with different scenarios in Smarter Round Rabin Algorithm those are

1. If the tasks in waiting list is less than or equal to three, then they proposed algorithm used Shorter Job First Algorithm and allocates the necessary time of until execution of each scheduled task.
2. If waiting list contains more than three tasks
   i) It calculate count of task in waiting list and count is pair number then time quantum is average burst time of tasks.
   ii) The task count in waiting list is an impair number then the time quantum is median tasks burst time.

5. Integrated Resource Scheduling Algorithm

Jia Ru (2013) [8] proposed an integrated resource scheduling algorithm for maximum utilization of resources and improves the performance. They designed this algorithm with combination of three algorithms such as task scheduling grouping algorithm, priority aware algorithm and shortest job first algorithm. An Integrated Resource Scheduling Algorithm is implemented in four phases those are as follows.

Initialize Phase

In this phase first they initialize the data for processing and available resources. A Gaussian distribution function is used to distribute the available tasks. The random task generation is used to follow the task processing requirements.

Prioritize Phase

In this phase allocates the resources based on prioritize. The highest transmission rate is gives first priority and scheduler sorts resources in descending order. As same as the resources which highest capacity is given

first priority. The scheduler sorts resources in descending order of processing capabilities. We can see the sorting function of transmission rate and processing capacity.

Sort(ResList, Descending Order of Transmission Rate)

Sort(ResList, Descending Order of Processing Capacity)

Grouping Phase

In the grouping phase constitutes the groups by adding the small tasks. The group is constitutes based resource requirement of each tasks. If a single task required high amount of resources that task considered as a one group. For example identify task MI > or < MIPS * Granularity Size. If task MI is larger than that task is grouped itself as one group. If task MI is lesser then group is formed with some of fine grained tasks[8].

SJF Scheduling Phase

In this phase they adopted SJF scheduling to reduce the task waiting time. In this phase they sorted grouped task based on ascending order of process requirement. As per this scheduling phase the shortest process requirement group is executed first and then second shortest group executed next. The rest of groups followed same process.

6. SMB Algorithm

Qiao Chu et al. (2017) [9] proposed a novel Stable Marriage Based(SMB) Algorithm. Using this algorithm they solved coupled placement problem of VMs in an integrated manner for storage and computing resources. To overcome the problem of 3-dimentional, they are used 3-dimentional stable matching relationship model among the VMs, compute nodes and storage nodes. The implementation of this algorithm they consider three main components such as applications, computing nodes and storage nodes.

A set of applications that are run on data center is $A = \{Ai : 1 \leq i \leq |A|\}$

A set of computing nodes on data center is $C = \{Cj : 1 \leq j \leq |C|\}$

A set of storage nodes on data center is $S = \{Sk : 1 \leq k \leq |S|\}$

They are identify some assumptions those are

The amount of computing resources are required for an application is denoted as ***Ai.Creq***

The amount of storage resources are required for an applications is denoted as ***Ai.Sreq***

Data transfer from storage to computing resource is denoted as ***Ai.Iorate*** Mbps

In this work they consider each applications contain a pair of compute node and storage node. As well as each storage node and computing node had a preference application list. The application which is required less capacity have high priority on storage node and also take less compute time. The preference list of computing node is depend on amount computing resource is required for a particular applications.

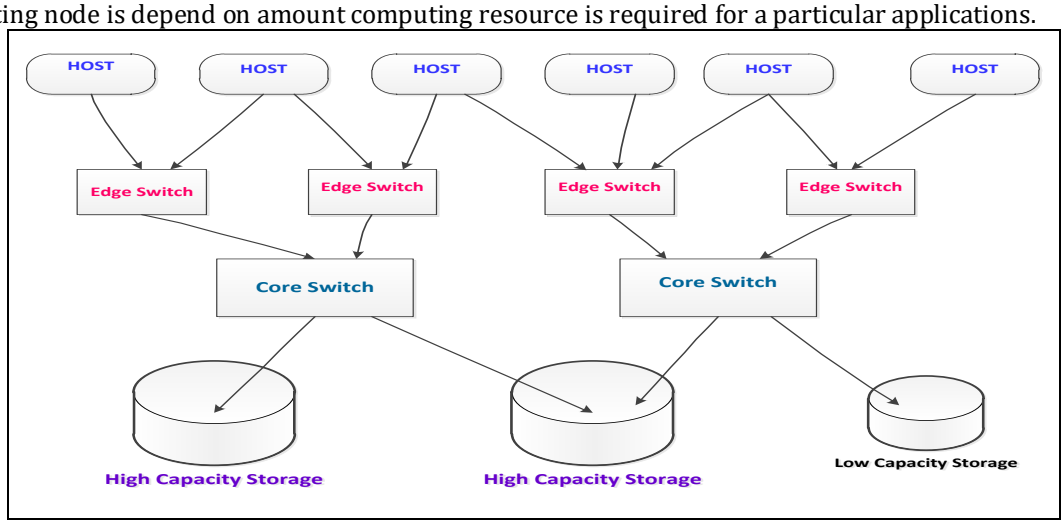

Fig 5. Architecture SMB Algorithm

The implementation of this algorithm is divided into three phases, such as Compute Phase, Storage Phase and Overall.

Compute Phase

In Compute phase of SMB algorithm followed as, each application is a pair of compute node and storage node. But in this phase they mainly focus on compute phase. Here many free applications are available. The compute node can pick up application from list of free applications. So the compute node can pick up less compute resource required applications. The remaining applications can reject by compute node. The compute node can also reject which application required high compute resource. The applications also

maintain the reject list of compute nodes. So this phase is best suitable for selection of applications by the compute nodes.

Storage Phase:

Actually the storage phase is similar to compute phase, here also the storage node can select the applications from the application preference list. Which applications is required less amount of storage can select by storage node first. In selection of application storage node consider order of preference , amount of memory required and capacity of storage node. Based on selection of application preference the VMs are allocated to application for processing[9].

Overall Phase:

In this overall phase, they combine the both phases and find best suitable matching to process the application efficiently.

7. Heuristic Algorithm

Qingshi Shao et al. (2018) [10] designed heuristic algorithm for resource scheduling in big data cloud applications. In this algorithm first they used admission control mechanism for job scheduling. The admission control mechanism operator is decide the acceptance of new arrival job and wait for execution or reject the job. Then the scheduler calculate the priority of all the jobs in the system. It also calculate the priority of running jobs and waiting jobs. The scheduler make a sorted queue based on current priority. All the jobs are placed in sorted queue and the scheduler picks up one by one and allocated the required resources. Here they need to determine priority of each new arrival job. So they proposed a novel heuristic, it maintain dynamic nature and execution of allocated jobs within specific amount of time. The main aim of this algorithm is reduce resource consumption is achieved.

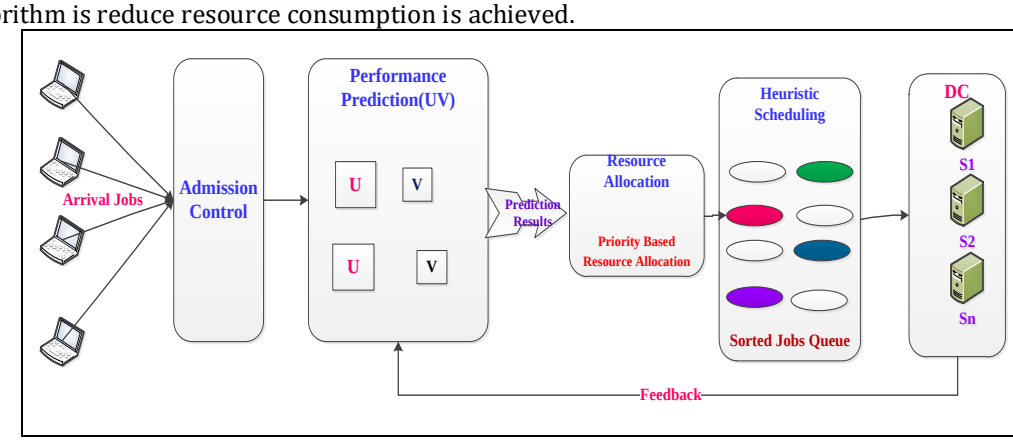

Fig 6. Working principle of heuristic scheduling algorithm

They used prediction application's runtime by integrate the prediction of every stage in application. In prediction application they used nine attributes as the columns in matrix, such as input-size, executor-number, executor-memory, executor-cores, stage-time.

Resource Allocation

Before creating the ultimate scheduling call for every job, an inexpensive and effective resource allocation strategy is necessary. With numerous runtime on totally different physical resource configurations, it's inadvisable to share resources proportional or assign several resources for all jobs shortly. Hence, associate applicable resource allocation strategy ought to mix the task combine, system load and job priority.

Summary

Table 2. Summary of resource scheduling algorithms reviewed

Author Name & Year	Algorithm Name	Advantages	Disadvantages
Yali Zhao et al. (2016)	Profit Optimization Resource Scheduling Algorithm,	Achieved profit maximization	This algorithm not applicable for large data sets
Saed Abed et al. (2018)	Task Scheduling Algorithm,	Increase resource utilization and reduce latency	Improvement of throughput is need
Hicham Gibet Tani et al.(2017)	Smarter Round Robin Scheduling Algorithm,	Reduce the execution time of big data	Real time deployment is not possible

		applications	
Jia Ru (2013)	Integrated Resource Scheduling Algorithm,	Achieved minimum waiting time	Not satisfy the real time clients
Qiao Chu et al. (2017)	SMB Algorithm,	Optimize the compute and storage resources	The energy consumption is high
Qingshi Shao et al. (2018)	Heuristic Algorithm,	Maximize the QoS constraints	This algorithm had some platform conditions

8. Conclusion

In this paper we made a survey of important resource scheduling algorithms found in the literature. Most of the resource scheduling algorithms used schedulers, task allocation, CPU time and VMs. In the same fashion, an optimization scheme can fall into different categories such as the resource utilization, the scheduled tasks, the big data applications. The researchers contributed towards resource scheduling using their simulation work with the tools of CloudSim, Hadoop Cluster, Apache Spark and Amazon EC2. In this paper explained different resource scheduling algorithms with different scenarios. Many such utilities were found in the literature. This paper focused on some of the important scheduling algorithms and their merits and demerits in solving the problem of resource scheduling and utilization. This research can be extended further by proposing a novel resource scheduling algorithm that can maximize performance analysis of big data cloud.

9. References

1. Huangke Chen et al. (2018). Big Data Processing Workflows Oriented Real-Time Scheduling Algorithm using Task-Duplication in Geo-Distributed Clouds. IEEE TRANSACTIONS ON BIG DATA. 1 (1),p 1-14.
2. Xiaomeng Su et al. (2013). Introduction to Big Data. NTNU. 1 (1), p1-11.
3. Lokesh Kumar Arya et al. (2014). Workflow Scheduling Algorithms in Cloud Environment - A Survey. IEEE. 1 (1),p 1-4.
4. Suman Sangwan et al. (2014). An Effective Approach on Scheduling Algorithm in Cloud Computing. IJCSMC. 3 (6),p 19-23.
5. Yali Zhao et al. (2016). SLA-Based Profit Optimization for Resource Management of Big Data Analytics-as-a-Service Platforms in Cloud Computing Environments. IEEE International Conference on Big Data. 1 (1),p 432-441.
6. Sa'ed Abed et al. (2018). Enhancement of Task Scheduling Technique of Big Data Cloud Computing. IEEE. 1 (1),p 1-6.
7. Hicham Gibet Tani et al. (2017). Smarter Round Robin Scheduling Algorithm for Cloud Computing and Big Data. Journal of Data Mining and Digital Humanities. 1 (1), p1-8.
8. Jia Ru et al. (2013). An Empirical Investigation on the Simulation of Priority and Shortest-Job-First Scheduling for Cloud-based Software Systems. IEEE. 1 (1), p78-87.
9. Qiao Chu et al. (2017). Joint Computing and Storage Resource Allocation Based on Stable Matching in Data Centers. IEEE. 1 (1), p207-212.
10. Qingshi Shao et al. (2018). A Market-oriented Heuristic Algorithm for Scheduling Parallel Applications in Big Data Service Platform. IEEE. 1 (1), p677-686.