Evento: XXV Jornada de Pesquisa
ODS: 9 - Indústria, Inovação e Infra-estrutura

UMA REVISÃO BIBLIOGRÁFICA DE MÉTRICAS QUE MEDEM O DESEMPENHO DOS PROCESSOS DE INTEGRAÇÃO DE DADOS[1]

A LITERATURE REVIEW OF METRICS TO MEASURE THE PERFORMANCE OF DATA INTEGRATION PROCESSES

Gabriela Gohlke Bley[2], Rafael Zancan Frantz[3]

[1] Pesquisa institucional desenvolvida no Departamento das Ciências Exatas e Engenharias-GCA/UNIJUÍ
[2] Aluno do Curso de Mestrado em Modelagem Matemática e Computacional, UNIJUÍ, e-mail: gbley22@gmail.com
[3] Professor permanente do PPG em Modelagem Matemática e Computacional, Orientador, UNIJUÍ, e-mail: rzfrantz@unijui.edu.br

Resumo

Com o constante crescimento das empresas, o aumento dos dados importantes em seus processos de negócios cresce juntamente. Dessa forma, as empresas recorrem a aplicativos de software para apoiar seus processos de negócios. É comum que uma empresa termine com uma ampla gama de aplicativos, que, na maioria dos casos, são desenvolvidos pelo departamento de TI da empresa ou adquiridos de empresas especializadas de software de terceiros. Nesta perspectiva, um processo de integração pode ser descrito como um fluxo de trabalho de tarefas interdependentes, vinculadas por canais de comunicação capazes de dessincronizar uma tarefa em relação a outra. Ou seja, as mensagens passam pelo fluxo de trabalho, encapsulando dados de / para os aplicativos integrados pelo processo. Portanto, um processo de integração é um software responsável pela coordenação externa de um conjunto de aplicativos, para que eles possam trocar dados e compartilhar funcionalidades para dar suporte a uma empresa. Nesta revisão, é relatada a analise de 8 artigos que tratam de plataformas de integração conhecidas no contexto atual. Foi realizada a metodologia de revisão e foi construída uma tabela listando as métricas utilizadas por cada autor e pelos artigos. Assim, foram encontradas 21 métricas nos 8 artigos analisados e constatou-se que o makespan é a métrica utilizada na maioria dos artigos.

Abstract

With the growth momentum of companies, the increase of important data in their business processes grows together. In this way, companies turn to software applications to support their business processes. It is common for a company to end up with a wide range of applications, which, in most cases, are developed by the company's IT department or acquired from specialized third-party software companies. In this perspective, an integration process can be described as a workflow of interdependent tasks, linked by communication channels that are capable of desynchronizing one task in relation to another. That is, the messages pass through the workflow, encapsulating data to/from the applications that are integrated by the process. Therefore, an integration process is a software responsible for the external coordination of a set of applications so that they can exchange data and share functionalities to support a company. In this review, it is reported the analyze of 8 articles that deal with integration platforms known in the current context. Was conducted a review methodology and a table was built listing the metrics used by each author and the articles. Thus, 21 metrics were found in the 8 analyzed articles and it was found that makespan is a metric used in most articles.

Evento: XXV Jornada de Pesquisa
ODS: 9 - Indústria, Inovação e Infra-estrutura

Palavras-chave: Fluxo De Trabalho, Processo de Integração , Desempenho, Métrica, Medida.
Keywords: Workflow, Integration Process, Performance, Metric, Measure.

1. INTRODUCTION

The efforts of companies in search of quality in its entirety have stimulated their search for technological alternatives to support and secure their business processes. Thus, companies acquire software in order to meet specific activities within the activities of business processes.
The software is usually quite diverse because the applications are developed with different technologies, that is, they run on different operating systems and/or use different data models. In this scenario, it is common for a business process to involve two or more applications and to make these applications work together, the company needs to integrate them so that such integration generates the least possible impact on their individual functioning.
The enterprise application integration area provides methodologies, techniques and tools that are used in the development of integration processes in order to connect the set of applications involved in the business process, making them collaborate with each other, that is, sharing their data or functionality. The integration processes are composed of tasks and have a workflow, where one task succeeds the other according to a dependency relationship in which the successor task depends on its predecessors. Integration platforms are specialized software tools for designing, implementing, monitoring and executing integration processes.
In this sense, this article aims to make a analyze between metrics that measure the performance of integration processes. Therefore, 8 articles were selected that deal with integration processes and some metrics currently used and then made a comparison between them and their results.
This work has the justification turned around for the future research for the author and the master's thesis, considering that will be in the area of the integration process. For this, the importance of this review, to know more about the metrics was utilized in the measure of the performance and be able to analyze them.
The results of this process was that knowledge about 21 metrics existents and found in the articles, but the more utilized metric in the same is the makespan. This, because the makespan, in the literature, is one of the most scheduling algorithms that have more focused on your optimizing. However, minimizing the total execution time also reduces the execution cost while mapping the tasks to the resources. This, shows the importance of the makespan in the integration process.
This article is organized as follows: in section of the research method, the review protocol is described, highlighting the methodology used, which is the literature review. After, the results are represented, subdivided into a definition of research questions, application of the review protocol, discussion of the selected articles, with a comparative table of the metrics that the articles contains and the last, description of the metrics. Lastly, presents the conclusions constructed in the review of the selected articles.

2. RESEARCH METHOD

This article has as methodology of scientific work for systematic literature review. This is used to believe that systematic literature review is a form of study that uses a well-defined methodology, identifying, analyzing and interpreting all the data used in respect of a research question, in an impartial and repeatable manner (KITCHENHAM et al., 2007).

Evento: XXV Jornada de Pesquisa
ODS: 9 - Indústria, Inovação e Infra-estrutura

Thus, 8 articles were analyzed that relate to addressing existing metrics to measure the performance of integration processes. Literature review is important because it recognizes the intellectual creation of other authors, employing an intellectual authority to the text.
Therefore, through the literature review, the author references and evaluates the knowledge produced in existing research, highlighting the most relevant points for the area in question, such as concepts, procedures, results, discussions and conclusions that he deems relevant to his work (PRODANOV et al., 2013).
In this way, it is know that the steps to be followed for the development of a systematic review are: elaboration of the research question, search in the literature, selection of articles, data extraction, evaluation of the methodological quality, description of the data (meta-analysis), evaluation of the quality of the samples and writing and publishing the results.

3. RESULTS

This section summarizes the results of the study.

3.1 DEFINITION OF RESEARCH QUESTIONS

The research questions are elaborated according to the general purpose of the systematic review study. This way, taking into consideration that the future research and master's thesis of the author will be in the area of the integration process, the objective of this work is get familiar and know more about the existing metrics to measure the integration process.
In this sense, a research question that addresses the objective of this work is: What are the measures capable of measuring the performance of the integration processes? To answer, articles focused on the data integration process were searched in Scopus that was a base of articles that published in journals.

3.2 APPLICATION OF THE REVIEW PROTOCOL

Knowing that the systematic review have steps to be followed, and that the first one is elaboration of the research question, that was explained in the previous section, the next step was the search in the literature. The same was occur through the platform Scopus with the search string: ({workflow} or {integration process} or {business process} or {integration solution}) and ({performance}) and ({metric} or {measure}). With this, was found 1,029 documents. So, was limited the subject area for computer science and document type to article, resulting in 167 articles. Of these, was selected 3 articles. The others articles was proposed to be studied by the master's advisor of the author of this work.
The selection of the articles developed tough the read of the abstract and the introduction of the works that was supposed most important in this view. After, was developed the studied of the complete work and the data extraction. With this datas was possible to observe that the quality of the samples was enough to writing and publishing good results and a good construction of the know for the author and for your research.

Evento: XXV Jornada de Pesquisa
ODS: 9 - Indústria, Inovação e Infra-estrutura

3.3 DISCUSSION OF THE SELECTED ARTICLES

In this section, the context and the problem involving the selected articles were presented. In this perspective, it is known that the demand for processing dynamically introduced tasks that require high computational power has increased dramatically in recent decades, as has research to face the many challenges it presents. In the work by Wangsom et al. (2019), evidences that execution of scientific workflows often involves computation and data intensive application. To improve the computation efficiency, a whole workflow usually is divided into multiple tasks. Then, those multiple tasks are allocated and processed over distributed systems such as cluster, grid, and currently, cloud computing. Over the years, Infrastructure-as-a-Service (IaaS) cloud becomes one of the majority computing resources for executing a large application because it offers unlimited, ondemand, scalable resources with the competitive performance at affordable price

From cloud users' perspective, cost and makespan, are probably the most common objectives for workflow scheduling as trade off between them are required. Moreover, due to their data-driven application by nature, scientific works usually generate a tremendous volume of data and transfer them between tasks during the execution. This has a significant impact on network utilization and energy consumption by network equipment in cloud data center. Hence, the cloud provider has the responsibility of creating a scheduling plan, not only to minimize the cost and makespan for satisfactory user requirements but also to reduce the data movement for decreasing energy consumption. This work proposes Peer-to-peer clustering technique for grouping dependent tasks in a workflow and allocating them to the same VM in order to reduce data movement (WANGSOM et al., 2019).

In another article by Sun et al. (2020), the authors proposed an aggregation measure factor-based scheduling algorithm (AFSA) for workflow applications, where time and cost parameters are simultaneously considered in a computational heterogeneous distributed environment. The objective of the proposed algorithm is to find the best scheduling of workflow tasks under task deadline and budget constraints that are predefined by a user or time and cost constraints that are predefined by a resource provider.

According Singh et al. (2019) a real world workflow application consists of a set of a large number of interdependent tasks. Such workflows can be represented as directed acyclic graphs (DAGs), which are executed on infrastructure as a service (IaaS) cloud to run the applications. An IaaS cloud is a deployment model used in cloud computing which provides computational resources to the users for executing their applications. Scheduling of workflows is the major concern for the cloud service provider (CSP) which furnishes IaaS cloud resources to its users on the basis of pay-as-you-go model. Many algorithms have been developed which consider various performance issues such as resource utilization, makespan, cost, fault tolerance and so on.

The algorithm presented in the article writed by Singh et al. (2019) is inspired from HCRO. However, the proposed method is shown to be more efficient with respect to minimization of energy consumption and the makespan. Besides, the proposed approach also considers energy conservation by using a DVS-enabled environment. It includes many aspects of real-time workflow scheduling such as CPU performance variability, VM boot time, VM shut down time, etc, which are not considered by HCRO. And, it incorporates a different criteria for including the new molecules in the current population, to enhance the scheduling results. The existing HCRO technique lacks all these features.

In the work by Singh et al. (2017) cloud computing is associated with a new era of technology that dynamically provisioned computing infrastructure and other services among end users. With

this approach, researchers become inclined toward the advantages offered for workflow applications. Current status of task scheduling in cloud computing is presented in various categories by many researchers. However, current cutting edge gave a broad review of scheduling algorithms in light of meta-heuristics techniques in context with distributed computing systems like grids and clouds. Presented the taxonomy and comparative review on these algorithms based on various scheduling objectives. Methodical analysis of task scheduling is presented based on swarm intelligence and bio-inspired techniques. For this, based on the analysis of various papers published between the years 2008 to 2016 in the area of scheduling in grids and cloud computing.

Besides, in one of the works writed by Anta et al. (2015) was consider a setting with a single machine prone to crashes and restarts that are being controlled by an adversary (modeling worst-case scenarios), and a scheduler that assigns injected jobs or tasks to be executed by the machine. These tasks arrive continuously and have different computational demands and hence size (or processing time).

In another paper by Abdi et al. (2014) PSO algorithm, genetic algorithm and modified PSO algorithm for task scheduling problem in Cloud computing environment are compared with each other. Simulation results show that even if three algorithms show acceptable revenue, but totally Modified PSO Algorithm performance is better than two other algorithms.

Cloud computing has emerged as the most popular distributed computing paradigm out of all others in the current scenario. It provides on-demand access to shared pool of resources in a self-service, dynamically scalable and metered manner with guaranteed Quality of service to users. To provide guaranteed Quality of Service (QoS) to users, it is necessary that Jobs should be efficiently mapped to given resources. If the desired performance is not achieved, the users will hesitate to pay (KALRA et al., 2015).

So, in the work by Kalra et al. (2015) was presented an extensive review of various scheduling algorithms based on five metaheuristic techniques namely Ant Colony Optimization (ACO), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), League Championship Algorithm (LCA) and BAT algorithm.

With the advancement of technology and emergence of grid and cloud computing, many large-scale scientific and engineering applications are usually constructed as workflows, which are similar to traditional parallel task graphs in structure, due to large amounts of interrelated computation and communication.

Huang et al. (2014) was proposed and evaluated new task-ranking and allocation methods for list-based workflow scheduling. The proposed approaches were evaluated with a series of simulation experiments and compared to well-known existing methods, heft and the lookahead variant of heft. The experimental results show that the approaches outperform existing methods significantly for some kinds of workflow structure properties and workload conditions.

So, the articles were analyzed and then a table was constructed in which the authors of the referenced articles are presented, relating to the metrics presented by them.

Evento: XXV Jornada de Pesquisa
ODS: 9 - Indústria, Inovação e Infra-estrutura

Table 1. Table listing the metrics adopted by each author.

Metrics	Studies							
	Wangsom et al.[1]	Sun et al.[2]	Singh et al.[3]	Singh et al.[4]	Anta et al.[5]	Abdi et al.[6]	Kalra et al.[7]	Huang et al.[8]
Cost	x		x	x			x	
Makespan	x	x	x	x		x	x	x
Data movement	x							
Budget				x				
Deadline				x				
Security				x				
Reliability				x				
Load balancing				x				
Resource utilization				x			x	
Rescheduling				x				
Energy Efficiency		x		x				
Completed load					x			
Pending load					x			
Latency					x			
Flow time							x	
Tardiness							x	
Waiting time							x	
Turnaround time							x	
Fairness							x	
Throughput							x	
Win								x

Analyzing the built table, it is clear that there are numerous metrics and that each of them takes into account different topics, always aiming to integrate the data in the best possible way, in a safe way. However, the metric that appears most in articles is the 'makespan'.

3.4 DESCRIPTION OF METRICS

In this section, the metrics that were found in each article analyzed, that was cited in the table 1. In this review will be discussed, bringing their definition, what is their powerlessness for the integration process to occur.

Makespan: Is described as the overall time required to execute whole workflow by considering the time when the tasks finished its execution and the time when it has been submitted. It can be defined as the duration from which the user submitted the workflow to the time it completes and gives results.

Latency: Is the largest time a task spends in the system, from the time of its arrival until it is fully executed. Latency, is also referred to as flowtime in scheduling.

Cost: It indicates the total amount the user needs to pay to service provider for resource utilization.
Data Movement: Data movement is occurred if the dependency edge exists and a parent task and a child task are assigned in different VM. Also, data movement in network equipment is one of the major portions that consumes a tremendous amount of energy.
Budget: Budget is basically constraint that is defined by the user to avail the services from the cloud service provider. Using the budget constraint, the scheduling decisions are made to minimize the total execution time of the workflow, and also it has to complete within the budget.
Deadline: Time critical applications are required to complete its execution within a certain time frame. Deadline constrained scheduling is designed for these applications to deliver results before the deadline. The deadline constrained scheduling also needs to consider monetary cost when it schedules tasks. Robust scheduling with deadline is much needed in time critical applications, and also it improves the application dependability.
Security: In cloud computing, resources are heterogeneous and distributed, so the term security becomes a serious issue. Providing data security and privacy in cloud environment is more complicated than the traditional systems because of virtualization and multi-tenancy feature. Security issues like data leakage and hyper visor vulnerabilities also considered in resource sharing and virtualization.
Reliability: Reliability is the probability of the tasks that run successfully and complete the execution of the workflow. Reliability means that all the resources to function well during the execution of an application. This objective should be considered as it reduces the chances of failures to complete the assigned workflow. The failure rate can be calculated so that mapping should be done to maximize the reliability and reduce the failure rates.
Load balancing: Mainly virtual machines are the processing elements in the cloud computing environments. In scheduling, there can be situation when more than one task is assigned to VMs to execute the tasks simultaneously. This leads to unbalancing of the loads over the VMs. The scheduler should be able to distribute the workload to the available resource in such a way that resources are not heavily loaded. Load balancing over the resource also improves the resource utilization and consequently improving the overall scheduling process performance.
Resource utilization: Increasing the resource utilization is beneficial to the service provider. To get maximum profit by renting the limited resources to the user in such a way that the resources are fully utilized.
Rescheduling: Rescheduling is mainly considered as an overhead to scheduling process because it leads to re-evaluating the schedule and the cost of data movement among the dependent tasks over the various machines. Not every task is selected for rescheduling as it increases the overall execution time and performances degrades.
Energy efficiency: CPU utilization and resource utilization directly effect on the energy consumed by a task. Energy consumption will be high when CPUs are not properly utilized because idle power is not effectively used. Sometimes it gives high in energy consumption due to heavy demand of the resources, and this may lower down the performance. The scheduling decisions are important to find the efficient sequence of tasks execution so as to reduce the energy consumption by the assigned resources.
Completed load: Is the aggregate size of all the tasks that have completed their execution successfully.
Pending load: Is the aggregate size of all the tasks that are in the queue waiting to be completed.
Flowtime: It is the sum of finishing times of all the tasks. To minimize the flowtime, tasks should be executed in ascending order of their processing time. Flowtime signifies the response time to the tasks submitted by users. Minimizing the value of flowtime means reducing the average response

Evento: XXV Jornada de Pesquisa
ODS: 9 - Indústria, Inovação e Infra-estrutura

time of the schedule.
Tardiness: This defines the time elapsed between the deadline and finishing time of a task i.e. it represents the delay in task execution. Tardiness should be zero for an optimal schedule.It is an important metric to measure the overall performance of the schedule with respect to meeting deadlines.
Waiting time: It is the difference between the execution start time and submission time of the task.
Turnaround time: This keeps track of how long it takes fora task to complete execution since its submission. It is the sum of waiting time and execution time of task.
Fairness: A desirable characteristic of scheduling process isfairness which requires that every task must get equal share of CPU time and no task should be starved.
Resource utilization: Another important criterion is maximization of resource utilization "i.e." keeping resources as busy as possible. This criterion is gaining significance as service providers want to earn maximum profit by renting limited number of resources.
Throughput: It is defined as the total number of jobs completing execution per unit time. Generally, a task-resource mapping schedule is optimized on the basis of single or multiple criterion.
Win: Used for the comparison of different scheduling algorithms. For any workflow, one or several of the evaluated algorithms would lead it to the shortest makespan. The win value of an algorithm means the percentage of the workflows that achieve the shortest makespan when applying the algorithm. From users' perspective, a higher win value represents that an algorithm has a more stable performance and might lead to higher satisfaction.
Thus, it can be seen that there are numerous known metrics and that each of them has its relevance in the process of data integration.

CONCLUSIONS

This article presents a systematic review that aim the analysis of the existing metrics was found in the selected articles, that measure the integration process. For this, was utilized the methodology of the systematic review and follow the protocol of you steps. In this way, was selected 8 works that involves this subject and studied about the data that their contributes for the area. The results of this study conclude that 21 metrics were found in the articles. All of them with their due importance, consider several factors to be analyzed, always aiming at data security.
Analyzing the results obtained, it is possible to verify that makespan is the most used metric by the authors of the articles, being cited in 7 of 8 selected articles. The workflow makespan, or execution time, is the time required to complete all tasks in the workflow of the integration process. The execution time has an impact on the performance of the integration process because it symbolizes the time it takes the data to be processed by all tasks in the process workflow. Thus, having knowledge of makespan helps configuration of the execution engine, being able to minimize the use of computational resources and meet quality of service requirements of the processes, hence the importance of this metric. Despite the fact that the workers consider makespan as a performance metric, it is known that the main objectives of its use are the reduction of computational resources, energy consumption and budgetary costs.
This work was development because, in the future, the master's thesis of the author of this article, will be in the area of the integration process. Thus, it is the extremely importance of the knowledge of the metrics that was used in the area, the same way, the evidence of the makespan has in this area.

Evento: XXV Jornada de Pesquisa
ODS: 9 - Indústria, Inovação e Infra-estrutura

REFERÊNCIAS

WANGSOM, P., LAVANGNANANDA, K., BOUVRY, P. Multi-objective scheduling for scientific workflows on cloud with peer-to-peer clustering. 11th International Conference on Knowledge and Smart Technology, 175-180. 2019

SUN, T., ZHANG, Y., XIONG, K., XIAO, C. Aggregation measure factor-based workflow application scheduling in heterogeneous environments. IEEE Access, 8, 89850-89865. 2020

SINGH, V., GUPTA, I., JANA, P. K. An energy efficient algorithm for workflow scheduling in IaaS cloud. Journal of Grid Computing. 2019

SINGH, P., DUTTA, M., AGGARWAL, N. A review of task scheduling based on metaheuristics approach in cloud computing. SpringerVerlag London. 2017

ANTA, A. F., GEORGIOU, C., KOWALSKI, D. R., ZAVOU, E. Competitive Analysis of Task Scheduling Algorithms on a Fault-Prone Machine and the Impact of Resource Augmentation. In: Springer International Publishing Switzerland 2015 F. Pop and M. Potop-Butucaru (Eds.): ARMS-CC 2015, LNCS 9438, 1–16. 2015

ABDI, S., MOTAMEDI, S. A., SHARIFIAN, S. Task Scheduling using Modified PSO Algorithm in Cloud Computing Environment. In: International Conference on Machine Learning, Electrical and Mechanical Engineering (ICMLEME'2014) Jan. 8-9, 2014

KALRA, M., SINGH, S. A review of metaheuristic scheduling techniques in cloud computing. In: Egyptian Informatics Journal, 16(3), 275-295. 2015

HUANG, K.C., TSAI, Y.L., LIU, H.C. Task ranking and allocation in listbased workflow scheduling on parallel computing platform. In: Springer Science+Business Media New York. 2014

KITCHENHAM, B., CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report. 2007

PRODANOV, C. C., FREITAS E. C. DE. Metodologia do trabalho científico [recurso eletrônico]: métodos e técnicas da pesquisa e do trabalho acadêmico – 2. ed. – Novo Hamburgo: Feevale. 2013

Parecer CEUA: 640.285