International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 3, March- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 6.505

A STUDY ON BIG DATA SCIENTIFIC WORKFLOWS IN THE CLOUD ENVIRONMENT: CHALLENGES AND FUTURE PERSPECTIVES

Parag Digambarrao Thakare
(Computer Science Engineering)
Dr. Rajeev Yadav
(Professor)
Glocal School of Technology and Computer Science

ABSTRACT

Several applications rely on scientific workflows to lessen the inherent complexity involved in scientific and business analytics. This involves mitigating the inherent complexity that is intrinsic to both types of analytics. Scientific workflow management systems have been shown to be beneficial in a variety of domains, such as social science, astronomy, neurology, and bioinformatics, to the extent that they are irreplaceable in their respective fields. This is supported by evidence. It is possible to define a workflow as a logical series of activities or tasks that are guided by rules and consist of data processing. The primary objective of this project is to carry out research on the topic of huge data scientific workflows in the cloud environment: challenges and future perspectives. The methodology employed in this study is qualitative research technique. The study identified SWMS (Scientific Workflow Management Systems) as cloud-based approaches for merging scientific workflows with big data analytics. This technique offers advantages such as scalability, flexibility, and ease of deployment. Several scheduling methods, systems, and structures have been developed to maximize these benefits.

Keywords: *Big Data Science; Cloud Environment; Scientific workflow; Big Data Analytics.*

INTRODUCTION

Several applications rely on scientific workflows to lessen the inherent complexity involved in scientific and business analytics. This involves mitigating the inherent complexity that is intrinsic to both types of analytics. Scientific workflow management systems have been shown to be beneficial in a variety of domains, such as social science, astronomy, neurology, and bioinformatics, to the extent that they are irreplaceable in their respective fields (Farley et al., 2018). This is supported by evidence. In the same way that traditional computing infrastructures have been demonstrated to be insufficient for dealing with the issues that are connected with big data analytics, data management and standard scientific workflows have also been proved to be inadequate. In addition, the scientific workflows that are currently in place

International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 3, March- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 6.505

are not capable of successfully addressing the issues that are brought about by the growing complexity and scale of analytical tasks (Hu et al., 2020).

It is possible to define a workflow as a logical series of activities or tasks that are guided by rules and consist of data processing. A workflow is a sequence of procedures that are applied to a process in order to automate that activity. Most workflows can be divided into two categories: (1) workflows used in the scientific field and (2) workflows used in business. It is possible to use a scientific workflow to the field of scientific computing in order to automate scientific experiments and processes (De Oliveira, Liu & Pacitti, 2019; Wen et al., 2020). The following section elaborates the past literatures related to this study.

LITERATURE STUDY:

Below is a table that provides a list of previous literatures that are relevant to this topic.

Table 1: Related works

AUTHORS AND YEARS	**METHODOLOGY**	**RESULTS AND FINDINGS**
Abualigah et al., (2021)	This paper introduced a hybrid Dragonfly Algorithm for intelligent Big Data task scheduling in IoT cloud computing applications.	The investigation, using a t-test, showed that MHDA (Mapping Dragonfly Algorithm for task scheduling) outperformed other algorithms in Big Data job scheduling due to its faster convergence and 17.12% improvement in results.
Aziza &Krichen (2020)	The goal is to determine the optimal allocation of process activities to available computing resources.	Experimental results indicated great efficiency of our suggested technique, making it suitable for cloud workflow scheduling. This study added a GA-based module into the WorkflowSim framework based on CloudSim.
Barika et al., (2019)	The orchestration requirements of these workflows were thoroughly covered in this study, along with the difficulties in meeting these criteria.	Big data processing is growing rapidly to generate insights that can impact enterprises, government policy, and research. This has improved communication, programming, and

		processing technologies including cloud computing, Hadoop, Spark, and Storm.
Hu et al., (2018)	This study presented a multi-objective scheduling (MOS) algorithm for scientific workflow in multicloud environments to decrease workflow makespan, cost, and dependability.	Numerous simulation experiments using real-world scientific workflow models show MOS algorithm's strong multi-objective performance increase.

RESEARCH GAP:

All workflow data cannot be kept during its execution lifecycle. Thus, superfluous data must be removed. There are various proposed workflows for running data-intensive tasks. Data-intensive operations need handling massive datasets. Parallelization is definitely the most effective way to streamline the execution process. Advantages of the Cloud include infinite resources for data-intensive workflows. So, this study focused on big data scientific workflows in the cloud.

METHODOLOGY:

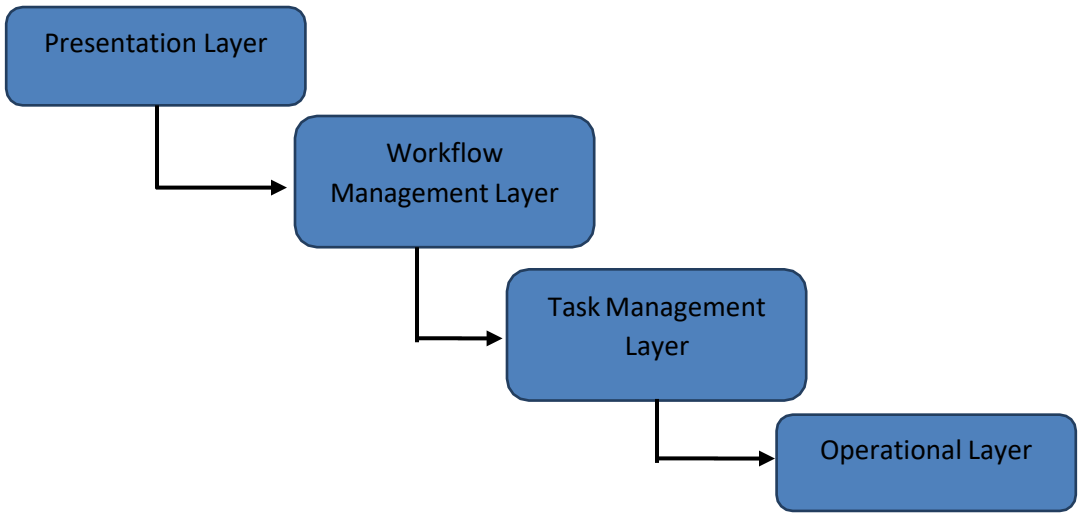
In this particular study, qualitative research methodology was utilized as the methodology of choice. To gain a better understanding of ideas, perspectives, or experiences, qualitative research entails the collection and examination of data that is not quantifiable in nature. Obtaining in-depth insights into a topic or coming up with new ideas for study are both possible use of this tool. Online databases such as Scopus, Google Scholar, and Web of Science, as well as online journals and publications, are the primary sources of information that were utilized in this investigation. This study only focused on the studies that were conducted after the year 2018.

RESULTS AND FINDINGS:

It is possible to visualize workflows through the use of a complicated graph that illustrates all of the processes that are running concurrently and need to be finished within a specific deadline. Access to data, processing of data, and display of data are all necessary components of any data analytics work. As a consequence of this, workflow tasks ought to guarantee that various aspects of the system are dealt with in an effective and efficient manner. It is possible to employ processes in cloud settings if they are refactored in accordance with the paradigm of cloud computing.

International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 3, March- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 6.505

The architecture was divided into four levels to integrate visualization, analytical engines, and data collection: workflow management, presentation layer, operational layer, and task management. This architecture was merely a foundation architecture, making it unsuitable for managing security issues. The diagram below illustrates the system architecture, which includes the following components. To integrate with Swift Workflow Management System, Eucalyptus and OpenNebula were used.



1. Figure 1: Architecture proposed by Hu et al., (2018)

The implementation of workflows for scientific research has typically been accomplished through the utilization of workstations, grids, clusters, and supercomputers. Scalability and computing complexity are the main restrictions and challenges that are linked with this technology. Scalability is the ability to increase the size of the system (De Oliveira et al., 2019; Hu et al., 2020). There have been a number of other problems that have been linked to the utilization of some resources. In the process of designing and utilizing scientific workflows for cloud-based big data analytics, there are a number of problems that need to be solved. In the natural world, the inputs and outputs of scientific operations are dispersed across several things. The data objects in question will be connected with a wide range of types, sizes, and complicated levels of sophistication. "Data deluge" is the term used to describe the massive amount of information that is produced by scientific computing. This information comes from a wide variety of sources, such as experiments, sensors, and networks (Wen et al., 2020).

International Journal of Research in IT and Management (IJRIM)
Available online at: http://euroasiapub.org
Vol. 12 Issue 3, March- 2022
ISSN(o): 2231-4334 | ISSN(p): 2349-6517 | Impact Factor: 6.505

Specifically, this is the key context in which research is being carried out concerning the deployment and execution of scientific workflows in a cloud environment inside a multisite environment. Taking this into consideration, a significant piece of work offers an architecture for the implementation of distributed systems for the management of scientific workflows. As a result of the deployment of the design that was described, there has been a significant reduction in costs.

The acronym SWMS stands for "scientific workflow management systems," and it refers to cloud-based technologies that integrate scientific workflows with big data analytics. Utilizing this strategy comes with a number of benefits, some of which include the capacity to scale, the capacity to be adaptable, and the ease with which it can be deployed. There have been many different scheduling algorithms, systems, and architectures that have been developed in order to take use of these benefits. So far, there has been a limited amount of study conducted on scientific workflow management systems. This article provides a comparison and illustration of the numerous frameworks and architectures that have been proposed and are now being utilized in order to make the most of the power that scientific processes possess. In addition to this, it investigates the feasibility of employing these systems for the purpose of doing analysis on substantial volumes of data.

CONCLUSION:

Finally, despite numerous systems addressing security concerns, the issue persists. Many research areas require clarification, including automating virtual cluster deployment for workflow applications and optimizing scheduling methods. Future research should include optimizing scheduling algorithms, completing workflows across different sites, and customizing systems to utilize edge computing capabilities.

DECLARATION:

I as an author of this paper / article, hereby declare that paper submitted by me for publication in the journal is completely my own genuine paper. if any issue regarding copyright/ patent/ other real author arises. the publisher will not be legally responsible. if any of such matters occur publisher may remove my content from the journal website/ updates. i have resubmitted this paper for the publication, for any publication matters or any information intentionally hidden by me or otherwise, i shall be legally responsible (complete declaration of the author at the last page of this paper/article).

REFERENCES

1. Barika, M., Garg, S., Zomaya, A. Y., Wang, L., Moorsel, A. V., & Ranjan, R. (2019). Orchestrating big data analysis workflows in the cloud: research challenges, survey, and future directions. *ACM Computing Surveys (CSUR)*, *52*(5), 1-41.
2. Abualigah, L., Diabat, A., & Elaziz, M. A. (2021). Intelligent workflow scheduling for Big Data applications in IoT cloud computing environments. *Cluster Computing*, *24*(4), 2957-2976.
3. Aziza, H., &Krichen, S. (2020). A hybrid genetic algorithm for scientific workflow scheduling in cloud environment. *Neural Computing and Applications*, *32*, 15263-15278.
4. Hu, H., Li, Z., Hu, H., Chen, J., Ge, J., Li, C., & Chang, V. (2018). Multi-objective scheduling for scientific workflow in multicloud environment. *Journal of Network and Computer Applications*, *114*, 108-122.
5. Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, *68*(8), 563-576.
6. Hu, Y., Wang, H., & Ma, W. (2020). Intelligent cloud workflow management and scheduling method for big data applications. *Journal of Cloud Computing*, *9*, 1-13.
7. De Oliveira, D. C., Liu, J., & Pacitti, E. (2019). *Data-intensive workflow management: for clouds and data-intensive and scalable computing environments*. Morgan & Claypool Publishers.
8. Wen, Y., Liu, J., Dou, W., Xu, X., Cao, B., & Chen, J. (2020). Scheduling workflows with privacy protection constraints for big data applications on cloud. *Future Generation Computer Systems*, *108*, 1084-1091.