DOI: 10.36558/rsc.v13i2.8282

Performance Modeling of Big Data Environments in the Private Cloud

Tarcisio J. Rolim Filho
Federal Rural University of Pernambuco
tarcisiorolim@ufrpe.br

Fernando A. A. Lins
Federal Rural University of Pernambuco
fernandoaires@ufrpe.br

Erica T. G. de Sousa
Federal Rural University of Pernambuco
erica.sousa@ufrpe.br

ABSTRACT

This work proposes the evaluation of the performance of big data environments in the private cloud through a methodology and a stochastic model the proposed methodology considers objective activities and performance modeling to assess Hadoop cluster performance in the private cloud. The stochastic model represents sending datasets to the Hadoop cluster with different configurations, and these infrastructures are represented through stochastic Petri nets. A case study based on the CloudStack platform and Hadoop cluster is adopted to demonstrate the feasibility of the methodology and the proposed model.

Keywords

Cloud Computing; Hadoop Cluster; Performance Evaluation; Dependability Evaluation; Stochastic Petri Nets; Reliability Block Diagram

1. INTRODUCTION

Cloud computing offers data storage services without the need for a dedicated infrastructure [1]. Therefore, both the infrastructure and the application interface are provided by a service provider. These can be adapted to the client's needs without having to participate in the installation, configuration, or maintenance of the product. In addition, resources such as virtual machines can be accessed by the client anywhere, anytime, also without the need to be involved in managing this infrastructure. Cloud computing provides several benefits such as reduced IT infrastructure cost for storing large data sets, scalability with rapid infrastructure expansion, and disaster recovery. There are generally three types of cloud deployments, private cloud, public cloud, and hybrid cloud [2].

Big data emerged to meet the demand for new techniques that allow the processing of information with high performance and availability [1]. Big Data tools make collecting, processing, and visualizing data simpler, more standardized, and more efficient. The benefit of cloud computing for the big data environment is the ability to provide a scalable and adaptable solution for large data sets and business analytics. Benchmarking the big data environment in the private cloud is important for planning how many computing resources the Hadoop cluster meets the requirements of a particular service.

This paper aims to evaluate the performance of Hadoop Cluster on private cloud infrastructures regarding a particular service. For this purpose, a performance model based on stochastic Petri nets and a methodology is proposed to evaluate the performance of Hadoop Cluster in private cloud infrastructures. The proposed methodology consists of activities of workload generation through the data set collected from social networks; performance measurement of Hadoop Cluster in cloud infrastructure; statistical analysis of performance metrics; representation of the Hadoop Cluster through the proposed performance model and analysis of new scenarios. The proposed performance model represents the workload and computational resources of Hadoop Cluster that are required for workload processing.

This work is organized as follows: Section 2 presents related work. Section 3 describes the main concepts for a better understanding of this work. Section 4 presents the methodology for evaluating the Hadoop Cluster in the private cloud. Section 5 details the proposed performance model. Section 6 presents the case study and Section 7 presents the conclusion and some future works.

2. BASIC CONCEPTS

This section presents the basic concepts for a better understanding of this work.

2.1 Stochastic Petri nets

Petri nets (PN) [20] is a family of formalisms very well suited for modeling several system types, since concurrency, synchronization, communication mechanisms as well as deterministic and probabilistic delays are naturally represented. Petri nets are a bipartite directed graph, in which places (represented by circles) denote local states and transitions (depicted as rectangles) represent actions. Arcs (directed edges) connect places to transitions and vise-versa.

This work adopts a particular extension, namely, Stochastic Petri Nets (SPN) [22], which allows the association of probabilistic delays to transitions using the exponential distribution or zero delays to immediate transitions (depicted as thin black rectangles). The respective state space can be translated into continuous-time Markov chains [21], and SPN also allows the adoption of simulation techniques for obtaining performance metrics (e.g.: response time, resource utilization, and throughput) and dependability metrics (e.g.: availability, reliability, and downtime), as an alternative to the Markov chain generation.

2.2 Phase Approximation Technique

Phase approximation technique can be applied for modeling non-exponential activities. A variety of performance and dependability activities can be constructed in SPN models by using throughput subnets and s-transitions, as depicted in Figures 1, 2 and 3. These throughput subnets and s-transitions represent polynomial-exponential functions, such as the Erlang, Hypoexponential and Hyperexponential distributions [15].

Measured data from a system (empirical distribution) with an average μ_D and a standard deviation σ_D must adjust their stochastic behavior through the phase approximation technique. The inverse of the variation coefficient of the measured figure (Equation (1)) allows the selection of which distribution matches it best. In this work, the adopted distribution for moment matching are the Erlang, Hypoexponential, and Hyperexponential distributions.

$$\frac{1}{CV} = (\frac{\mu_D}{\sigma_D}) \tag{1}$$

When the inverse of the variation coefficient is a whole number and different from one, the empirical figure should be characterized by an Erlang distribution, represented in SPN by a sequence of exponential transitions whose length is calculated by Equation (2). The rate of each exponential transition is calculated by Equation (3). The Petri Net model depicted in Figure 1 represents an Erlang distribution.

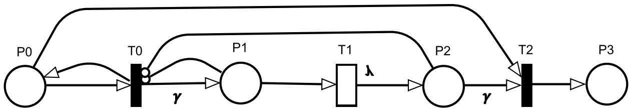

Figure 1: Erlang Distribution Net

$$\gamma = (\frac{\mu}{\sigma})^2 \tag{2}$$

$$\lambda = (\frac{\gamma}{\mu}) \tag{3}$$

When the inverse of the variation coefficient is a number larger than one (but not an integer), the empirical figure is represented by a hypoexponential distribution which is represented by a SPN composed of a sequence whose length is calculated by Equation (4). The transition rates of exponential transitions are calculated by Equations (5) and (6) where the respective average time (expected values) assigned to the exponential transitions are calculated by the Equations (7) and (8). The model presented in Figure 2 is a net that depicts a hypoexponential distribution.

$$(\frac{\mu}{\sigma})^2 - 1 \leq \gamma < (\frac{\mu}{\sigma})^2 \tag{4}$$

$$\lambda_1 = (\frac{1}{\mu_1}) \tag{5}$$

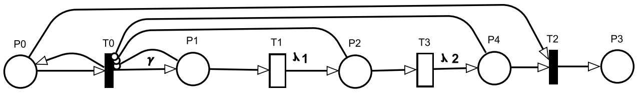

Figure 2: Hypoexponential Distribution Net

$$\lambda_2 = (\frac{1}{\mu_2}) \tag{6}$$

$$\mu_1 = \mu \mp \frac{\sqrt{\gamma(\gamma+1)\sigma^2 - \gamma\mu^2}}{\gamma+1} \tag{7}$$

$$\mu_2 = \gamma\mu \pm \frac{\sqrt{\gamma(\gamma+1)\sigma^2 - \gamma\mu^2}}{\gamma+1} \tag{8}$$

When the inverse of the variation coefficient is a number smaller than one, the empirical figure should be represented by an hyperexponential distribution. The exponential transition rate should be calculated by Equation (9) and the weights of immediate transitions are calculated by Equations (10) and (11). The Petri Net model that represents this hyperexponential distribution is depicted in Figure 3.

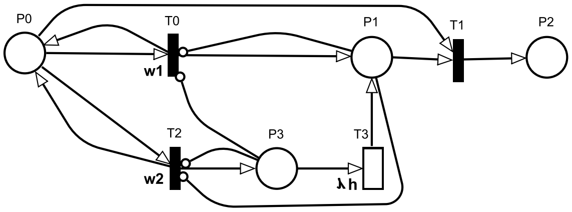

Figure 3: HyperExponencial Distribution Net

$$\lambda_h = (\frac{2\mu}{\mu^2 + \sigma^2}) \tag{9}$$

$$w_1 = (\frac{2\mu^2}{\mu^2 + \sigma^2}) \tag{10}$$

$$w_2 = 1 - w_1 \tag{11}$$

3. RELATED WORK

This section presents the related work to the performance evaluation of big data environments in cloud infrastructures. Ohnaga et al. [16] discuss the performance of a Hadoop application program running on such hybrid clouds. A performance model to estimate the execution time of a Hadoop application program running on a hybrid cloud is presented.

This paper [17] proposes a performance analysis model for Big Data Applications. The main goal of this work is to fill the gap that exists between the quantitative (numerical) representation of quality concepts of software engineering and the measurement of performance of Big Data Applications. For this, it is proposed the use of statistical methods to establish relationships between extracted performance measures from Big Data Applications, Cloud Computing platforms, and software engineering quality concepts.

This paper [18] aims to evaluate the performance of MongoDB database using the Yahoo Cloud Serving Benchmark tool, with 3 cloud application profiles from Google and Microsoft.

Rista et al. [19] tackle the problem of improving network performance in container based on cloud instances to create a viable alternative to run network-intensive Hadoop applications. The proposed approach consists of deploying link aggregation via the IEEE 802.3ad standard to increase the available bandwidth and using Linux Container to create a Hadoop cluster. In order to evaluate the efficiency of the proposed approach and the overhead added by the container, a set of experiments was conducted to measure throughput, latency, bandwidth utilization, and completion times.

We can observe that only one related work presents a methodology for the performance evaluation of big data applications in cloud computing; all papers present performance measurement results, and one paper presents a performance model. This work contains as contributions a methodology, measurement, and model for the performance evaluation of Hadoop Cluster in Private Cloud.

4. METHODOLOGY FOR PERFORMANCE EVALUATION OF HADOOP CLUSTER IN PRIVATE CLOUD

This section presents the proposed methodology to evaluate the performance of the Hadoop Cluster [5] configured in the private cloud. The proposed methodology is composed of the following activities, as shown in Figure 4: understanding and configuration of the big data environment in private cloud; planning of experiments of big data environment in private cloud; workload generation of social network data; performance modeling of big data environments in private cloud; performance measurement the of big data environments in private cloud; statistical analysis of performance metrics of big data environments in private cloud; performance modeling of big data environments in private cloud; performance model refinement of big data environments in the private cloud; performance metrics mapping of big data environment in the private cloud; performance model validation of big data environments in private cloud; and analysis of new scenarios of big data environment in private cloud.

4.1 UNDERSTANDING AND CONFIGURATION OF THE BIG DATA ENVIRONMENT IN PRIVATE CLOUD

. In order to assess the performance of the big data environment configured on private cloud infrastructures, it is necessary to understand the purpose of the analysis. Afterward, based on the service requirements that will be offered, it will be chosen the cloud platform and the big data application. This activity also considers the identification of metrics for performance evaluation of the big data application in the private cloud. The chosen private cloud platform must be configured considering virtual machines with different service offerings. The main platforms of private clouds that can be adopted are Apache Cloudstack, Apache OpenStack, and Eucalyptus [6]. Likewise, the Hadoop Cluster [5] must be configured considering the different amounts of data nodes and master nodes.

4.2 PLANNING OF EXPERIMENTS OF THE BIG DATA ENVIRONMENT IN PRIVATE CLOUD

. In regards to planning the experiments [15] to evaluate the performance of the Hadoop Cluster in the private cloud infrastructure, the computing capacity offered by the private cloud will be identified to the instantiation of virtual machines. This allows the establishment of factors and their levels for the planning of experiments [7]. The Hadoop Cluster is composed of the master nodes and data nodes [5] which are configured in the private cloud. These components are configured in different service offerings of cloud infrastructure. The number of master nodes and data nodes depends on the supply of services provided by the private cloud. Therefore, the service offering and the number of data nodes can be considered factors of the design of experiments, and their variations are considered the levels of these factors. The types of design experiments that can occur are simple, complete, and factorial [7].

4.3 WORKLOAD GENERATION OF SOCIAL NETWORK DATA

. The activity of workload generation provides data from social networks that will be analyzed in the big data environment configured in the private cloud. The data set can be captured from social networks through software tools, such as RStudio [8] and a data capture algorithm for analysis purposes of statistical calculations and data analysis of social networks. This data set can be processed and analyzed by software, such as MapReduce and Apache Spark.

4.4 PERFORMANCE MEASUREMENT OF BIG DATA ENVIRONMENTS IN PRIVATE CLOUD

. This activity provides the measurement of data regarding selected performance metrics, considering a particular configuration of the Hadoop Cluster [5] in the private cloud and a workload. The measurement of selected metrics takes place in each experiment generated in the design of experiments, for at least 30 times [9]. Metrics such as execution time (sec) and resource utilization can be adopted for benchmarking big data environments in the private cloud. On this activity measurement and sampling intervals for the collection of performance metrics are also defined. Tools like SYSSTAT and PERFMON [10] measure the performance metrics like resource utilization and response time [11]. Moreover, the private cloud environment and virtual machines must be reset after each experiment to avoid changing the metrics results. At the time of storing the log with the performance metrics data, it is checked if there is any inconsistency with the measured data. If there is, new performance measurements will be held. Otherwise, it will be performed statistical analysis of performance metrics data.

4.5 STATISTICAL ANALYSIS OF PERFORMANCE METRICS OF BIG DATA ENVIRONMENT IN PRIVATE CLOUD

. This activity aims at the statistical analysis of the performance metrics. The result of this analysis is the calculation of the averages and standard deviations of the performance measures adopted for each scenario configured according to the factors and levels defined in the design of experiments. In addition, the analysis of the existence of outliers is carried out, which may have been caused by minor errors, such as disturbances in the measurement environment.

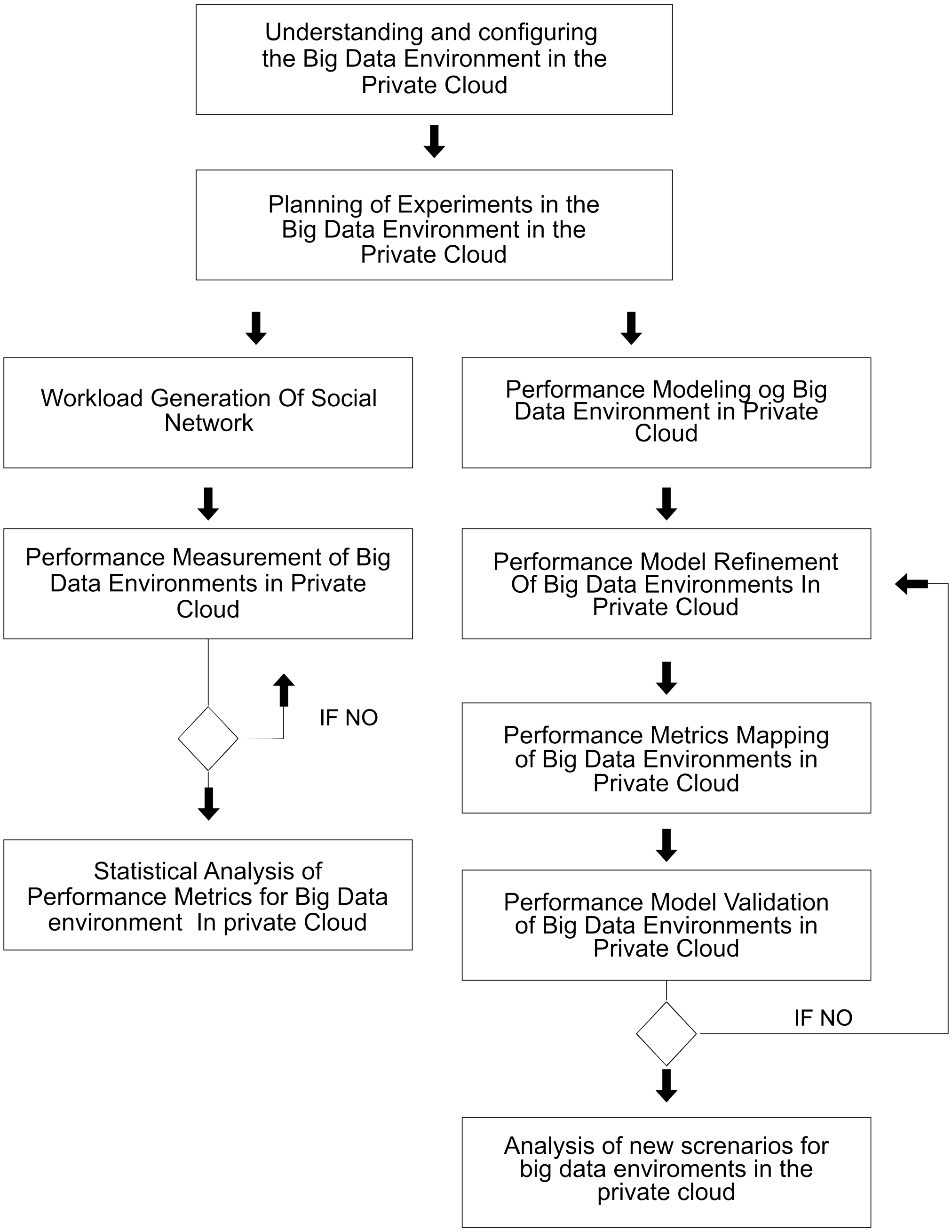
Understanding and configuring the Big Data Environment in the Private Cloud
Planning of Experiments in the Big Data Environment in the Private Cloud
Workload Generation Of Social Network
Performance Modeling og Big Data Environment in Private Cloud
Performance Measurement of Big Data Environments in Private Cloud
Performance Model Refinement Of Big Data Environments In Private Cloud
IF NO
Performance Metrics Mapping of Big Data Environments in Private Cloud
Statistical Analysis of Performance Metrics for Big Data environment In private Cloud
Performance Model Validation of Big Data Environments in Private Cloud
IF NO
Analysis of new screnarios for big data enviroments in the private cloud

Figure 4: Methodology for Evaluating Hadoop Cluster Performance in Private Cloud

4.6 PERFORMANCE MODELING OF BIG DATA ENVIRONMENTS IN PRIVATE CLOUD.

. The proposed performance model considers the stochastic Petri nets formalism [3] for the performance evaluation of the Hadoop cluster [5] configured in the private cloud. For this purpose, the workload and the Hadoop cluster components must be modeled, such as the master node and data nodes.

4.7 PERFORMANCE MODEL REFINEMENT OF BIG DATA ENVIRONMENTS IN PRIVATE CLOUD.

. The model refinement is performed from the performance metrics collected at the measurement activity. These metrics are adopted to generate the performance model parameters. The phase approximation technique provides the selection of the expolinomial probability distribution and the numerical parameters of this probability distribution that best represent the metrics that were collected for performance evaluation of the big data environment in the private cloud [4].

4.8 PERFORMANCE METRICS MAPPING OF BIG DATA ENVIRONMENT IN PRIVATE CLOUD

. Elements of the proposed performance model are used to represent the selected performance metrics. The purpose of this activity is to represent the set of performance criteria for big data applications in private clouds through elements of stochastic Petri nets [3], since this mathematical formalism was adopted to design the refined model.

4.9 PERFORMANCE MODEL VALIDATION OF BIG DATA ENVIRONMENT IN PRIVATE CLOUD

. Performance model validation allows the comparison of metric results obtained through the refined performance model and the metrics collected in the measurement activity. Comparing the results of these metrics should be equivalent to an error of acceptable accuracy. If the value of this error is greater than 10%, it will be necessary to refine the performance model [7]. If the accuracy error is equal to or less than 10%, it will be carried out the analysis of new scenarios. The t-pair test can also be used to quantitatively assess the model of refined performance.

4.10 ANALYSIS OF NEW SCENARIOS OF BIG DATA ENVIRONMENTAL IN PRIVATE CLOUD

. This activity aims to analyze new scenarios with different workload amounts, several configurations of virtual machines, and quantities of data nodes. For this activity, the validated performance model is adopted to perform the analysis of selected metrics.

5. PERFORMANCE MODEL OF HADOOP CLUSTER IN PRIVATE CLOUD

This section presents the performance model of Hadoop Cluster [5] configured in the private cloud, as shown in Figure 5.

This model is based on stochastic Petri net [3] and is composed of Workload and Hadoop Cluster subnets [5]. In the Workload subnet, there is the marking (N) assigned to the place Client that defines the workload that will be sent to the Hadoop Cluster, where the number of markings is proportional to the size of the data set. The sending time of Workload is associated with timed transition SEND (ST). After triggering the timed transition SEND, the request is sent to be serviced by the Hadoop cluster. After triggering the immediate transition Send Data SET, the request is sent to the master node. The time associated with the timed transition Process Master Node (PMN) represents the time needed to coordinate the processing activities. The time associated with the timed transition Process Data Set (PD) represents the time required for data nodes in the Hadoop cluster to process the data set. After this time, the resource is released. The N associated with the place DATANODE represents the number of data nodes that compose the Hadoop Cluster. The markings of the places TOTALMEMORY and TOTALPROCESSOR represent Hadoop cluster memory and processor capacities, where the capacity of each data node is added to represent the total cluster capacity. The data set is processed using the resources of the data nodes. Once the data set is processed, the resources of the processor and memory of the virtual machine are released [12]. Table 1 presents the parameters of the Hadoop cluster in private cloud performance model.

Table 1: PARAMETERS OF THE PROPOSED PERFORMANCE MODEL

Transition	Time	Type	Weight	Priority	Concurrency
SEND	ST	EXP	-	-	SS
SEND DATASET	-	IMM	1	1	SS
PROCESS MASTER NODE	PMN	EXP	-	-	SS
PROCESS DATASET	PD	EXP	-	-	SS
RESOURCE RELEASED	-	IMM	1	1	SS
REQUEST MEMORY	-	IMM	1	1	SS
PROCESS RESOURCE	-	IMM	1	1	SS
REQUEST PROCESSOR	-	IMM	1	1	SS

According to Table 2, the UP metric represents processor utilization. E{#MASTERNODE} represents the average number of markings in the place MASTERNODE, E{#MASTERNODE PROCESSING} represents the average number of markings in place MASTERNODE PROCESSING and E{#PROCESSING} represents the average number of markings in place PROCESSING. The place TOTAL PROCESSOR represents the sum of the processing core of the data nodes that constitute the Hadoop Cluster. The UM metric represents memory utilization. E{#TOTAL MEMORY} represents the average number of tokens in the place TOTAL MEMORY. The parameter TOTAL MEMORY represents the total memory capacity of the Hadoop Cluster deployed in the private cloud, in GB.

Table 2: PERFORMANCE METRICS OF THE PROPOSED PERFORMANCE MODEL

METRIC	EXPRESSION
UP	((E{#MASTERNODE}) + (E{#MASTERNODE PROCESSING}) + (E{#PROCESSING})) X 100 / (TOTAL PROCESSOR))
UM	((TOTAL MEMORY) - (E{#TOTAL MEMORY})) X 100 / (TOTAL MEMORY)

6. CASE STUDY

The case study presented in this section aims to assess the performance of the Hadoop Cluster [5] configured on the private cloud infrastructure. In this study, the maximum workload supported by Hadoop Cluster in the private cloud is also evaluated. In this environment, map reduce is adopted to analyze a data set composed of phrases from users of the social network Twitter that made posts adopt-

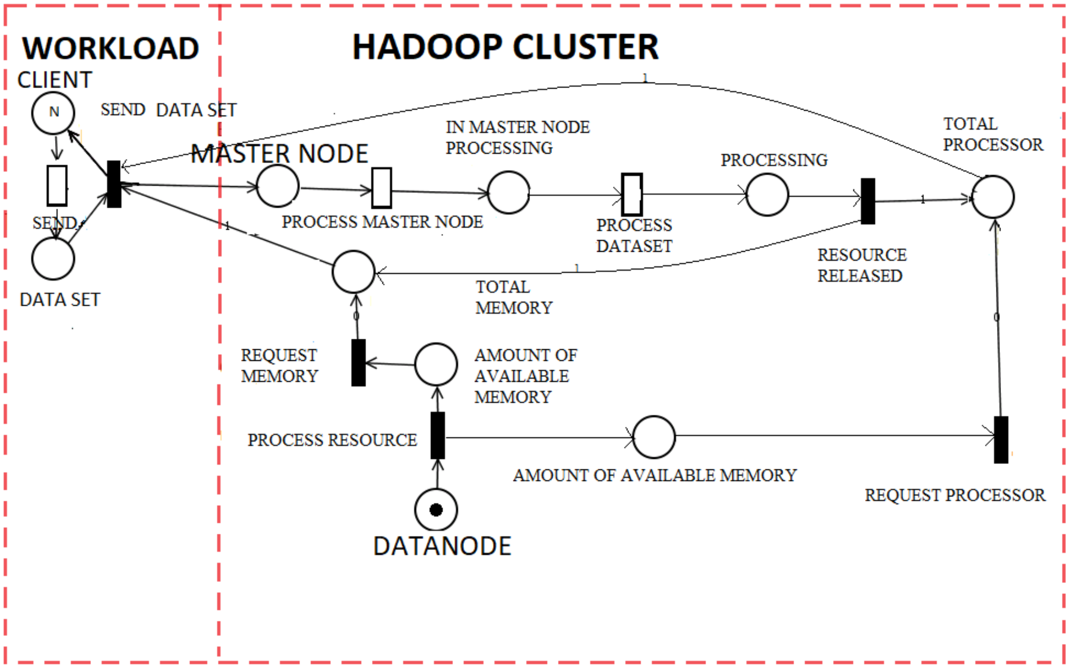

Figure 5: Hadoop cluster in private cloud performance model

ing words that denote symptoms of depression. According to the Pan American Organization of Health [13], mental disorders as a whole are responsible for approximately 13% of diseases in the world, and more than 300 million people, of all ages are affected by such disorders. Depression is a functionally and socially disabling illness, which holds different symptoms and periods for each individual.

6.1 UNDERSTANDING AND CONFIGURATION OF THE BIG DATA ENVIRONMENT IN PRIVATE CLOUD

. The adopted environment for performance evaluation is a private cloud consisting of 7 computers, with an operating system without an interface to reduce resource consumption and with the Cloud Stack platform. The machines present the configuration as shown in Table 3.

Table 3: PRIVATE CLOUD INFRASTRUCTURE SERVICE OFFERING.

Resources	Configuration
Memory	4GB | 8GB
Processor	I3 | I5
HD	1TB
Hypervisor	KVM
S.O.	CentoS 7

5 machines have an I5 processor and 8GB memory, and 2 machines have I3 processor and 4GB memory.

6.2 PLANNING OF EXPERIMENTS OF THE BIG DATA ENVIRONMENT IN PRIVATE CLOUD

. In these planning of experiments, some factors with different levels were adopted. The service offering defines private cloud capacity for data nodes configured on virtual machines. Table 4 shows the service offering adopted for the planning of experiments.

Table 4: SERVICE OFFERING OF PRIVATE CLOUD INFRASTRUCTURE.

Service Offerings	Configuration
Small	Mem:4GB,Proc: 6 Cores,Storage:1TB
Medium	Mem:6GB,Proc: 6 Cores,Storage:1TB
Large	Mem:8GB,Proc: 8 Cores,Storage:1TB

The customized factors and levels selected in Table 5 for the planning of experiments were adopted according to the processing and memory capacity of the private cloud.

To generate the workload, an analysis of the feelings of Twitter users who made posts using words adopted by people with symptoms of depression was performed.

6.3 WORKLOAD GENERATION OF SOCIAL NETWORK DATA

. In order to generate workload, it was made sentiment

Table 5: FACTORS AND LEVELS OF PLANNING OF EXPERIMENTS

Service Offering	Workload	Data nodes number
Small	3GB	3
Medium	4GB	4
Large	5GB	5

analysis through Twitter users who made posts using words adopted by people with symptoms of depression was performed. Between the period of December 5th to 19th, 2021, 5GB of data were collected, which were converted into a dataset. The tool used in this data collection process was RStudio [8] which is an application that can capture Twitter data with the execution of a script, through a package called twitteR [8]. The capture of posts on this social network was carried out according to surveyed sentiment analysis parameters in the medical literature and on the website of the Ministry of Health of the federal government of Brazil [14].

6.4 PERFORMANCE MEASUREMENT OF BIG DATA ENVIRONMENTS IN PRIVATE CLOUD

. In this activity, the experiments were replicated 30 times, and the results of the averages of the processor utilization and memory utilization metrics were computed. Table 6 presents the results of the average execution time (second), the average metric of the processor utilization and memory utilization of data nodes configured in virtual machines instantiated in the private cloud, and the results of processor utilization and memory utilization metrics calculated through the proposed performance model. In this table, S means scenarios, ET represents the average running time, %UPMed means the average of the processor utilization measured, %UPMod represents the processor utilization calculated through the model, %UMMed denotes the average of the memory utilization measured, %UMMod means memory utilization obtained through the model, SF represents service offering, W represents the Workload and DN denotes the number of data nodes.

6.5 STATISTICAL ANALYSIS OF PERFORMANCE METRICS OF BIG DATA ENVIRONMENTS IN PRIVATE CLOUD

. In each experiment, the processor utilization and the memory utilization metrics were collected. Then, there was the removal of outliers from the performance metrics. Afterward, the averages of these metrics were calculated.

6.6 PERFORMANCE METRICS MAPPING OF BIG DATA ENVIRONMENTS IN PRIVATE CLOUD

. The performance metrics contemplated at the mapping are the processor utilization and memory utilization of data nodes configured in virtual machines of the private cloud, once that these metrics provide the Hadoop cluster planning [12].

6.7 PERFORMANCE MODEL REFINEMENT OF BIG DATA ENVIRONMENTS IN PRIVATE CLOUD

. For this work, the refining of the performance model is developed by the approximation technique of phases, which calculates the first and second moments of the empirical probability distribution of the execution times metric. In this work, the empirical probability distribution of the execution times metric was normal. The refinement of the performance model occurred with the parameterization of this model considering the hypoexponential probability distribution that represents the execution time metric collected in 30 replications of each of the 45 experiments planned according to Table 6.

6.8 PERFORMANCE MODEL VALIDATION OF BIG DATA ENVIRONMENTS IN PRIVATE CLOUD

. The validation of the proposed performance model was carried out using the paired T-test, which compares the average difference between two independent samples, in this case, the processor utilization and memory utilization metrics measured in the 45 experiments and calculated through the performance model. Considering a significance level of 5%, the paired T-test generated a confidence interval of (-1.073;1.057) for the memory utilization metric and (-3.20;1.45) for the processor utilization metric. As the confidence interval contains 0, there is no statistical evidence to reject the hypothesis of equivalence between the measured execution times and those obtained from the performance model.

6.9 ANALYSIS OF NEW SCENARIOS OF BIG DATA ENVIRONMENTAL IN PRIVATE CLOUD

. The new scenarios consider the maximum workload supported by the Hadoop cluster [5] with 3, 4, and 5 data nodes configured in the private cloud infrastructure with the large service offering, according to Table 7. These new scenarios were simulated using the proposed performance model. Thus, the performance model's Workload subnet represented the workloads 10GB, 15GB, and 25 GB through the N parameter of the place Client.

Table 8 presents the processor utilization and memory utilization of data nodes configured on the virtual machines of the private cloud. It can be observed that higher load intensities are applied, regarding the big data environment with 10, 15, and 25 GB data nodes. Again, in this table, S means scenarios, %UPMed means the average of processor utilization measured, %UPMod represents the processor utilization taken through the model,%UMMed denotes the average of memory utilization measured, %UMMod means the memory utilization obtained through the model, SF represents Service offering, W denotes the Workload and DN is the number of data nodes.

The Workload and data node number present an impact on the variation of the processor and memory utilization metrics in the new scenarios. In these scenarios, the processor utilization metric has not reached the level of saturation. However, the memory utilization metric reached the saturation level in all scenarios, with utilization values higher than 99%, indicating a need for resizing this feature to avoid loss of performance at the data set processing. These results of the memory utilization metric have already been observed in other studies [23] [24] [25].

7. CONCLUSION

The contributions of this work were the proposal of a methodology and a stochastic model to the performance evaluation of big data environments in the private cloud. This work provides the performance evaluation of the Hadoop cluster in the private cloud through the measurement and statistical analysis of the metrics execution time, processor utilization, and memory utilization. In order to generate the workload, sentiment analysis was performed on

Table 6: PROCESSOR AND MEMORY UTILIZATION METRICS

S	ET (sec)	UP Med (%)	UP Mod (%)	UM Med (%)	UM Mod (%)	SF	W	DN
1	0.03	19.13	22.66	61.50	60.02	S	3GB	3
2	0.03	20.53	21.32	59.44	56.60	S	3GB	4
3	0.04	21.83	20.56	51.86	47.33	S	3GB	5
4	0.03	29.83	23.24	23.24	19.41	M	3GB	3
5	0.04	31.15	22.39	22.39	16.06	M	3GB	4
6	0.04	32.33	21.49	21.49	9.33	M	3GB	5
7	0.03	34.62	22.74	20.34	21.17	L	3GB	3
8	0.03	35.96	21.43	20.12	20.36	L	3GB	4
9	0.04	37.27	20.65	20.01	15.84	L	3GB	5
10	0.04	42.74	21.76	22.78	20.98	S	4GB	3
11	0.04	44.12	20.89	21.62	16.51	S	4GB	4
12	0.04	45.52	20.01	20.76	12.08	S	4GB	5
13	0.04	41.47	22.76	22.45	14.42	M	4GB	3
14	0.04	43.26	21.54	21.47	9.79	M	4GB	4
15	0.04	32.37	20.85	20.87	9.49	M	4GB	5
16	0.03	38.56	23.15	21.87	14.26	L	4GB	3
17	0.04	39.81	22.12	21.04	9.25	L	4GB	4
18	0.04	40.23	21.74	20.65	7.43	L	4GB	5
19	0.04	39.34	23.08	23.88	15.57	S	5GB	3
20	0.05	40.71	22.89	22.64	10.87	S	5GB	4
21	0.05	41.42	21.67	21.68	9.12	S	5GB	5
22	0.04	40.57	22.06	21.88	16.12	M	5GB	3
23	0.04	41.14	21.59	21.10	11.34	M	5GB	4
24	0.05	42.68	21.13	20.54	10.45	M	5GB	5
25	0.04	40.96	22.67	22.30	17.09	L	5GB	3
26	0.05	41.80	21.46	21.77	12.65	L	5GB	4
27	0.05	43.45	21.21	21.23	11.76	L	5GB	5

Table 7: PRIVATE CLOUD INFRASTRUCTURE SERVICE OFFERING

Service offering	Workload (GB)	Data node number
Large	10	3,4,5
Large	15	3,4,5
Large	25	3,4,5

Twitter users' posts with words that indicated symptoms of depression. The proposed performance model was based on stochastic Petri nets and provided the evaluation of the processor utilization and memory utilization of the Hadoop cluster in the private cloud, considering different service offerings, workloads, and the number of data nodes. The case study, based on the CloudStack platform and the Hadoop cluster, considered data sets of different sizes formed from the sentiment analysis of Twitter users' posts. The results of the processor and memory utilization metrics obtained through the validated performance model showed that the memory resource saturates when the large service offering was adopted for setting 3, 4, or 5 data nodes, with workloads of 10GB, 15GB, and 25GB. These results demonstrated that workload is the biggest influencing factor in the memory utilization of big data applications in the private cloud.

The performance measurement of the Hadoop cluster in the private cloud considered a maximum size of the dataset for the generation of the Big Data workload of 5GB and a maximum number of data nodes of 5 due to restrictions related to the computational capacity of the private cloud.

Table 8: PROCESSOR UTILIZATION AND MEMORY UTILIZATION METRICS OF NEW SCENARIOS

S	UP Mod (%)	UM Mod (%)	SF	W	DN
1	72.68	99.11	L	10GB	3
2	75.98	99.36	L	10GB	4
3	77.43	99.55	L	10GB	5
4	72.62	99.03	L	15GB	3
5	76.03	99.43	L	15GB	4
6	77.40	99.51	L	15GB	5
7	72.72	99.16	L	25GB	3
8	75.97	99.35	L	25GB	4
9	77.42	99.54	L	25GB	5

Execution time, processor utilization, and memory utilization metrics were adopted to evaluate the performance of the Hadoop cluster in the private cloud, but the evaluation of this environment can consider metrics such as availability and performability. In future work, we intend to evaluate the impact of availability on the performance of the Hadoop cluster configured in the private cloud.

8. REFERENCES

[1] A. KHALIFA, AND M. ELTOWEISSY, *Collaborative autonomic resource management system for mobile cloud computing.*, IARIA, Proceedings of the Fourth International Conference on Cloud Computing, GRIDs, and Virtualiza-

tion, pages 115–121, 2013.

[2] Y. TAMURA AND S. YAMADA, *Software reliability analysis considering the fault detection trends for big data on cloud computing.*, Springer-Verlag Berlin Heidelberg 2015 , Lecture Notes in Electrical Engineering, pages 1021–1030, 2015.

[3] M. K. MOLLOY, *Performance Evaluation Using Stochastic Petri Nets.*, Performance Evaluation Using Stochastic Petri Nets. V. C-31, n. 9, p. 913-17, 1982. 2015.

[4] S. T. YEE AND J. A. VENTURA, *Phase-type approximation of stochastic petri nets for analysis of manufacturing systems.*, IEEE Transactions on Robotics and Automation., v. 16, n. 3, p. 318–322, Jun. 2000.

[5] APACHE SOFTWARE FOUNDATION., Disponível: *http://hadoop.apache.org/*, 2022.

[6] B. D. MARTINO, G. CRETELLA AND A. ESPOSITO, *Cloud Portability and Interoperability.*, in Issues and Current Trends, Springer International Publishing, 1ª ed, 2015.

[7] D. A. MENASCE, L. W. DOWDY AND V. A. F. ALMEIDA, *Performance by design.* in computer capacity planning by example, Prentice Hall, 1ª ed., 2004.

[8] RSTUDIO., Disponível: *https://rstudio.com/*, 2022.

[9] HP PERFORMANCE ENGINEERING BEST PRACTICES SERIES., Disponível: *https://softwaresupport. softwaregrp.com/*, 2022.

[10] WINDOWS. START WINDOWS RELIABILITY AND PERFORMANCE MONITOR IN A SPECIFIC STANDALONE MODE. [S.L.], Disponível: *https://docs.microsoft.com/en-us/windows-server/administration/windows-commands/perfmon*, 2022.

[11] G. JUVE, B. TOVAR, R. F. D. SILVA , D. KRÓL, D. THAIN, E. DEELMAN, W. ALLCOCK AND M. LIVNY, *Practical resource monitoring for robust high throughput computing.*, In: IEEE, 2015 IEEE International Conference on Cluster Computing, p. 650–657, 2015.

[12] T. ZHAO, S. WANG, J. ZUO, X. DUAN, AND X. WANG, *Performance Evaluation Of Smart Meters Based On Grey Relational Analysis.*, In: IEEE, 2018 10th International Conference on Intelligent Man-Machine Systems and Cybernetics (IHMSC), 2018.

[13] OPAS. ORGANIZAÇÃO PAN-AMERICANA DE SAÚDE. FOLHA INFORMATIVA- DEPRESSÃO., Disponível: *http:// saúde-folha-informativa/depressão.2018.*, 2022.

[14] DEPRESSÃO: CAUSAS, SINTOMAS, TRATAMENTOS, DIAGNÓSTICO E PREVENÇÃO, 2022., Disponível: *https://saude.gov.br/saude-de-a-z/depressao.*, 2022.

[15] A. A. DESROCHERS AND AL-R. Y. JAAR, *Applications of Petri Nets in Manufacturing Systems: Modeling, Control, and Performance Analysis.*, IEEE Press, 1995.

[16] H. OHNAGA, K. AIDA AND O. ABDUL-RAHMAN, *Performance of Hadoop Application on Hybrid Cloud.*, 2015 International Conference on Cloud Computing Research and Innovation (ICCCRI), pp. 130-138, 2015.

[17] L. E. VILLALPANDO, A. APRIL AND A. ABRAN, *Performance analysis model for big data applications in cloud computing.*, J Cloud Comp 3, 19, 2014.

[18] F. NADEEM, *A Unified Framework for User-Preferred Multi-Level Ranking of Cloud Computing Services Based on Usability and Quality of Service Evaluation.*, Access IEEE, vol. 8, pp. 180054-180066, 2020.

[19] C. RISTA, D. GRIEBLER, C. A. F. MARON AND L. G. FERNANDES, *Improving the Network Performance of a Container-Based Cloud Environment for Hadoop Systems.*, 2017 International Conference on High Performance Computing Simulation (HPCS), 2017, pp. 619-626, 2017.

[20] MURATA, T., *Petri Nets: Properties, Analysis and Applications, Proc. IEEE, number 4, volume 77, pages 541-580*, 1989.

[21] TRIVEDI, K., *Probability and Statistics with Reliability, Queueing and Computer Science Applications, Wiley Interscience Publication, Edition 2*, 2002.

[22] GERMAN, R., *Performance Analysis of Communication Systems with Non-Markovian Stochastic Petri Nets, John Wiley & Sons, Inc. New York, NY, USA*, 2000.

[23] DU, J. AND HUANG, Q., *Implementation of power monitoring data cloud platform based on Hadoop, IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2018.

[24] FOTACHE, M. AND CLUCI, MARIUS-IULIAN, *Big Data Performance in Private Clouds. Some Initial Findings on Apache Spark Clusters Deployed in OpenStack, 20th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, 2021.

[25] ASHAYER, A., YASROBI, S., THOMAS, S. AND TABRIZI, N., *Performance Analysis of Hadoop Cluster for User Behavior Analysis, IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2018.