Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 266-288
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Performance and Availability Evaluation of Big Data Environments in the Private Cloud

Tarcísio Rolim, Erica Sousa

Federal Rural University of Pernambuco, Recife, Brazil
Email: tarcisiorolim@gmail.com, erica.sousa@ufrpe.br

How to cite this paper: Rolim, T. and Sousa, E. (2024) Performance and Availability Evaluation of Big Data Environments in the Private Cloud. *Journal of Computer and Communications*, **12**, 266-288.
https://doi.org/10.4236/jcc.2024.1212015

Received: November 5, 2024
Accepted: December 28, 2024
Published: December 31, 2024

Abstract

Cloud computing allows scalability at a lower cost for data analytics in a big data environment. This paradigm considers the dimensioning of resources to process different volumes of data, minimizing the response time of big data. This work proposes a performance and availability evaluation of big data environments in the private cloud through a methodology and stochastic and combinatorial models considering performance metrics such as execution times, processor utilization, memory utilization, and availability. The proposed methodology considers objective activities, performance, and availability modeling to evaluate the private cloud environment. A performance model based on stochastic Petrinets is adopted to evaluate the big data environment on the private cloud. Reliability block diagram models are adopted to evaluate the availability of big environment data in the private cloud. Two case studies based on the CloudStack platform and Hadoop cluster are adopted to demonstrate the viability of the proposed methodologies and models. Case Study 1 evaluated the performance metrics of the Hadoop cluster in the private cloud, considering different service offerings, workloads, and the number of data sets. The sentiment analysis technique is used in tweets from users with symptoms of depression to generate the analyzed datasets. Case Study 2 evaluated the availability of big data environments in the private cloud.

Keywords

Cloud Computing, Big Data, Hadoop Cluster, Performance Evaluation, Availability Evaluation, Reliability Block Diagram, Stochastic Petri Nets

1. Introduction

Cloud computing is a technology that allows you to distribute your computing services and access them online without installing programs. With this, its services

can be accessed remotely, anywhere in the world, and at any time desired. The distribution of services is carried out through a service platform via the Internet, with a price defined according to use.

Therefore, cloud computing can provide faster innovation, flexible resources, and economies of scale. This service offers fast access to flexible IT resources, action on a global scale, increased productivity, better performance, and greater security [1]. Big data refers to large data sets collected, stored, and analyzed using advanced computing technologies to extract meaningful insights. This data can be structured, semi-structured, or unstructured and often comes from social networks, IoT sensors, mobile devices, and financial transactions.

The motivation for using big data lies in its ability to provide valuable information and insights for companies to make informed and strategic decisions. For example, companies can use big data to better understand their customers, optimize their production processes, and manage financial risk.

Cloud computing can help improve big data performance and availability in several ways. First, the cloud offers on-demand scalability, allowing companies to easily scale up or down their processing and storage resources as big data demands. Companies do not need to invest in their hardware and infrastructure to handle large data sets.

The state-of-the-art presents works that propose the performance evaluation of big data environments in cloud computing. The authors [2] presented the performance evaluation of Deep Learning (DL) applications in the big data environment. The evaluation measured the impact of parallel and distributed processing on Ten-sorFlow DL. The results showed better performance in distributed processing, with an acceleration of up to 8x and a loss of less than 5% precision.

The authors [3] evaluated the performance of the Hadoop cluster by measuring the execution time metric. As an extension of this work, they presented the performance evaluation of Hadoop, Spark, and Flink through the Terasort benchmark. The results showed that replacing Hadoop with Spark or Flink can significantly improve data processing performance.

Other works present the performance and availability evaluation of big data environments in cloud computing. The authors [4] propose RBD and SPN models for evaluating the capacity-oriented availability of nodes in a Eucalyptus private cloud. These calculations indicate how many virtual machines will be available on a node. The work also presents the resources lost due to failures and repairs. In the same study, [4] demonstrates that RBD models are used to calculate metrics such as stationary availability and annual downtime. The SPN model is used to calculate capacity-oriented availability.

The work [5] proposed models of performance, availability, cost, and a mechanism for evaluating design space for cloud infrastructure with a support (cluster system integrated to the front end) to a video service for a cluster subsystem on an independent physical machine. It was used as the cloud manager in the

Eucalyptus environment. In this performance study, JMeter was used for load generation and validation of the proposed model, and to calculate the availability, the MTTF and MTTR of the hardware and system were calculated. Operating system availability increases significantly to 98% in the scenario with an independent physical cluster for 99% in the scenario where we have a Front-end and a cluster together on the same physical medium.

Some works also show the availability assessment of cloud environments. The proposed work [6] combines low-level RBD models and high-level RBD models conforming to cloud computing components to availability evaluation.

In this way, the research problem that motivates this paper is described in the following question: How can the performance and availability of big data environments in the private cloud be evaluated?

This work proposes a strategy based on a methodology and Stochastic Petri Nets (SPNs) and Reliability Block Diagrams (RBDs) models for evaluating the performance and availability of big data environments in the private cloud.

This paper is divided into 7 Sections, which will be briefly highlighted in this section. Section 2 presents the basic concepts of the proposed work. Section 3 presents the methodology for evaluating the performance of big data environments in private clouds. Section 4 presents the methodology for evaluating the availability of big data environments in private clouds. Section 5 introduces performance and availability models. Section 6 presents the case study to evaluate the cluster's performance configured in the private cloud infrastructure according to the proposed methodology and models. This section also evaluates the availability of the Hadoop cluster configured in the private cloud infrastructure according to the method for evaluating the availability of big data environments in the private cloud, the hierarchical modeling strategy, and the proposed RBD models. Section 7 presents the conclusions.

2. Basic Concepts

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

2.1. Stochastic Petri Nets

Petri nets (PN) [7] is a family of formalisms very well suited for modeling several system types since concurrency, synchronization, communication mechanisms, and 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 vice versa.

This work adopts a particular extension, namely, Stochastic Petri Nets (SPN) [8], 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 [9], 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

The phase approximation technique can be applied to model non-exponential activities. A variety of performance and dependability activities can be constructed in SPN models by using throughput subnets and s-transitions. These throughput subnets and s-transitions represent polynomial-exponential functions, such as the Erlang, Hypoexponential, and Hyperexponential distributions [10].

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 distributions for moment matching are the Erlang, Hypoexponential, and Hyperexponential distributions.

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

2.3. Reliability Block Diagram

Reliability Block Diagram (RBD) [11] is a combinatorial model initially proposed for calculating systems' related reliability and availability metrics using intuitive block diagrams. The blocks (e.g.: components) are usually arranged using the following composition mechanisms: series, parallel, bridge, *k*-out-of-*n* blocks, or a combination of previous compositions.

Availability is the probability of a system being in a functioning condition. It considers the alternation of operational and nonoperating states. Steady-state availability (A) is commonly adopted, and the following equations are also taken into account: $A = uptime/(uptime + downtime)$, or $A = MTTF/(MTTF + MTTR)$. MTTF is the mean time to failure, and MTTR is the mean time to repair.

3. Methodology for Performance Evaluation and Availability Evaluation of Big Data Environment in Private Cloud

This section presents two methodologies: the first for evaluating the performance of big data environments in cloud computing and the second for evaluating their availability.

3.1. Methodology for Performance Evaluation of Big Data Environment in Private Cloud

The proposed methodology is composed of 10 activities, as shown in **Figure 1**. They are understanding, objectives, and configuring the big data environment in the private cloud, planning experiments in the big data environment in the private cloud, generating the workload of social network data, performance modeling of

big data environments in the private cloud, performance measurement of big data environments in the private cloud, statistical analysis of performance metrics of big data environments in the private cloud, refinement of the performance model of big data environments in the private cloud, mapping performance metrics of big data environments in the private cloud, validation of the performance model of big data environments in the private cloud, and analysis of new scenarios of big data environments in the private cloud.

1) Understanding, Objectives, and Configuration of the Big Data Environment in the Private Cloud—To evaluate the performance of the big data environment configured in private cloud infrastructures, it is necessary to understand the requirements of the configured services. Then, based on these requirements, the cloud platform and the big data application were chosen. This activity also considered the identification of metrics for evaluating the performance 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 private cloud platforms that can be adopted are Apache Cloudstack [12], Apache OpenStack [13], and Eucalyptus [13]. Likewise, the big data environment Hadoop cluster [14] could be configured considering the different numbers of data nodes and master nodes.

2) Planning Experiments in the big data environment in the private cloud—When planning experiments, the computational capacity offered by the private cloud to instantiate the virtual machines will be identified. This allows the establishment of factors and their levels. The Hadoop cluster is made up of Master nodes and Data nodes that are configured in the private cloud. These components are configured into different cloud infrastructure service offerings. The number of Master nodes and data nodes depends on the service offered by the private cloud. In this way, the service offer and the number of data nodes can be considered factors in the design of experiments and their variations in the levels of these factors. In this way, small, medium, and large would be the levels of the service offer, and 2, 4, and 6 would be the levels of the number of data nodes.

3) Generating the Workload of Social Network data—The workload generation activity 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 [15] and a data capture algorithm for statistical analysis purposes. This dataset can be processed and analyzed by MapReduce and Apache Spark [14].

4) Performance Measurement of Big Data Environments in the Private Cloud—This activity measures selected performance metrics, considering a given configuration of the big data environment in the private cloud and a workload. The selected metrics are measured in each designed experiment at least 30 times. Metrics such as execution time (sec) and resource utilization (%) can be adopted to evaluate the performance of big data environments in the private cloud. In this activity, measurement and sampling intervals for metric collections are also defined.

Tools like SYSSTAT [16] and PERFMON [17] collect performance metrics such as resource utilization and response time.

5) Statistical Analysis of Performance Metrics of Big Data Environments in the Private Cloud—This activity aims to analyze the statistical data of the chosen performance metrics. The result of this analysis is the calculation of the averages and standard deviations of the performance metrics adopted, referring to each scenario configured according to the factors and levels defined in the experiment planning. In addition, an analysis of outliers that may have been caused by minor errors, such as disturbances in the measurement environment, is carried out using the Minitab tool [18].

6) Performance Modeling of Big Data Environments in the Private Cloud—Cloud applications can be modeled by stochastic Petri nets, in which client/server [19] models communicate through cloud computing network interfaces. Important metrics for the quality and management of cloud applications can be extracted from models based on stochastic Petri nets [1].

The proposed performance modeling is based on SPN [16] and considers the performance evaluation of the Hadoop cluster configured in the private cloud. The proposed performance modeling considers Hadoop clusters composed of master and data nodes. The master node coordinates and manages cluster resources, and the data nodes analyze data sets that different sources, such as social networks, meteorological data, and health data, can generate.

7) Refinement of the Performance Model of Big Data Environments in the Private Cloud—The model refinement is based on the metrics collected in the measurement activity. The phased approximation technique provides the selection of the hypoexponomyal probability distribution and the numerical parameters of this probability distribution that best represent the metrics collected to evaluate the performance of the big data environment in the private cloud [20].

8) Mapping Performance Metrics of Big Data Environments in the Private Cloud—This activity aims to represent the set of performance criteria for big data applications in private clouds through elements of stochastic Petri nets since this mathematical formalism was adopted to design the refined proposed model.

9) Validation of the Performance Model of Big Data Environments in the Private Cloud—The validation of the performance model allows the comparison of the performance metrics obtained through the refined model and collected metrics. Comparing the results of these metrics must be equivalent to an acceptable accuracy error. If the value of this error is greater than 10%, it will be necessary to refine the performance model again [21]. If the precision error is equal to or less than 10%, the analysis of new scenarios will be performed. The refined performance model can quantitatively evaluate the paired test [21].

10) Analysis of New Scenarios of Big Data Environments in the Private Cloud—This activity aims to analyze new scenarios with different workload volumes, various virtual machine configurations, and numbers of data nodes. This activity adopts the validated performance model to analyze processor and memory

utilization metrics.

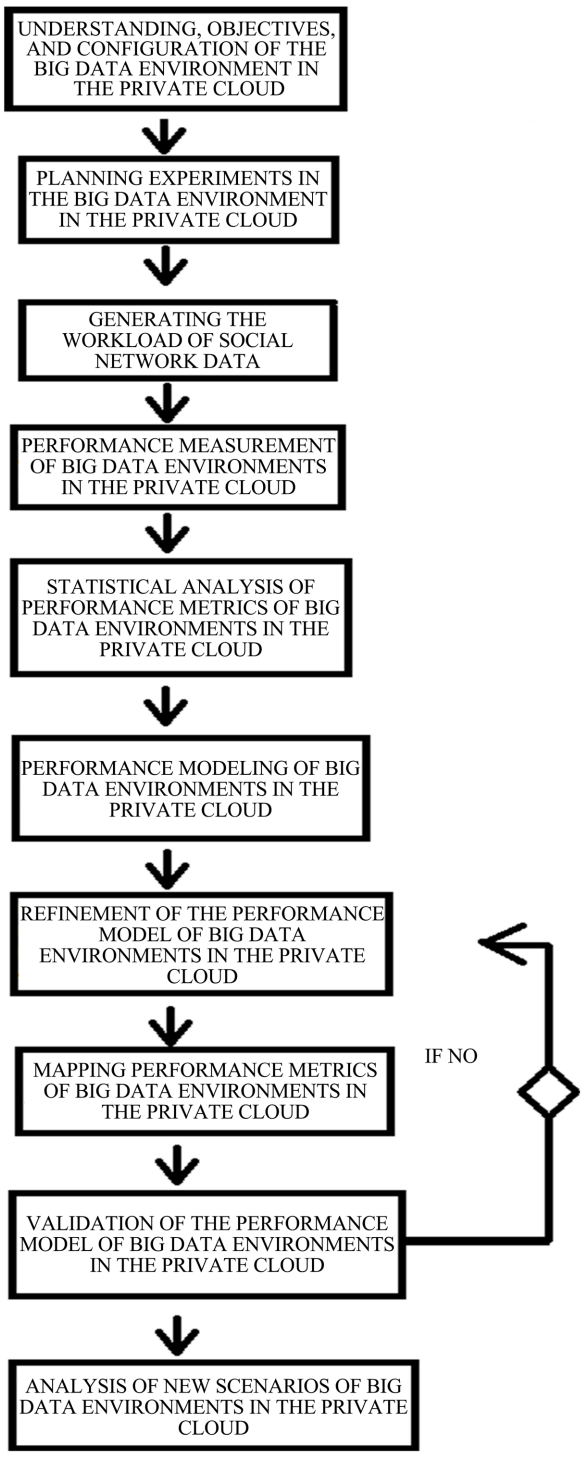

Figure 1. Methodology for performance evaluation of big data environments in the private.

3.2. Methodology for Assessing Big Data Environment Availability in Private Cloud

This section presents the proposed methodology (Figure 2) to evaluate the

availability of big data environments configured in the private cloud. This methodology used models based on RBD to represent the system availability metric. Modeling through RBD was chosen due to the possibility of representing the interdependence of its system components.

The proposed methodology consists of four activities: understanding and configuring the big data environment in the private cloud, generation of availability models, parameterization of availability models, and scenario analysis of big data environments in the private cloud.

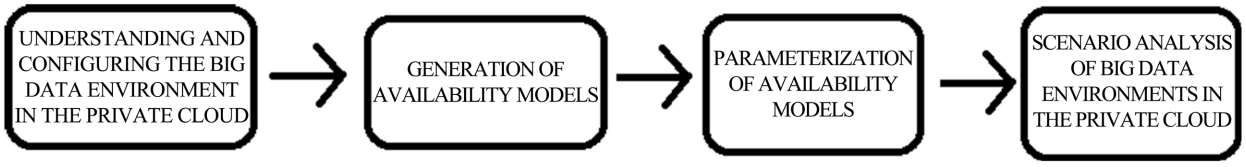

Figure 2. Methodology for assessing the availability of big data environments in the private cloud.

1) Understanding and Configuring the Big Data Environment in the Private Cloud—This activity evaluates the availability of big data environments in cloud infrastructures, considering the different services offered. This activity also considers the choice of cloud platform and big data environment. The main private cloud platforms that can be adopted are Apache Cloudstack [12], Apache OpenStack [13], and Eucalyptus [22]. Likewise, the Hadoop cluster [14] can be configured considering different amounts of data nodes and master nodes.

2) Generation of Availability Models—This activity aims to provide models to evaluate the availability of big data environments in the private cloud. Thus, a hierarchical modeling strategy was proposed to combine models based on RBD. In this modeling strategy, low-level models are combined to calculate the parameters of the high-level model. These models represent the private cloud infrastructure.

3) Parameterization of Availability Models—In this activity, the RBD models adopted for evaluating the availability of the big data environments in the private cloud will be parameterized. The parameters used are MTTF (Mean Time to Failure) and MTTR (Mean Time to Repair) are the components of the private cloud and big data environment. The values of these parameters can vary according to the adopted scenario.

4) Scenarios Analysis of Big Data Environments in the Private Cloud—This activity aims to analyze the availability of new scenarios considering the configuration of big data environments in the private cloud, such as equipment with different MTTF values.

4. Models

This section presents an SPN model and RBD models to evaluate the performance and availability of big data environments in the private cloud, respectively.

4.1. Performance Model

This section presents the proposed SPN model (Figure 3) for evaluating the performance of big data environments in private cloud infrastructures. The Workload and Hadoop Cluster sub-models of the proposed performance model represent the client's requests and the big data environment, respectively.

The Workload subnet represents the sending of user requests to the Hadoop cluster configured in the private cloud. The tag (NC) assigned to the Client place defines the workload that will be sent to the Hadoop cluster, where the number of tokens is proportional to the size of the data set. After triggering the immediate Send transition, the request is sent to be serviced by the Hadoop cluster. After triggering the immediate Data SET transition, the request is sent to the master node.

The Hadoop Cluster subnet represents the processing and storage infrastructure of the private cloud used for configuring the master node and data nodes. The master node has the function of coordinating processing activities, which are represented by the immediate transition PROCESS MASTER NODE. The time (TT) associated with the timed transition PROCESS DATA SET represents the time required for the data nodes of the Hadoop cluster to process the dataset. After this time, processing and memory resources are released. The marking (ND) associated with the place of the DATANODE represents the amount of data nodes that make up the Hadoop Cluster. The (MT) and (PT) tokens of the TOTAL MEMORY and TOTAL PROCESSOR places represent the memory and processor capacities of the Hadoop cluster, where the capacity of each data node is summed to represent the total capacity of the cluster. The dataset is processed using the data nodes' resources. Once the dataset is processed, the virtual machine's processor and memory resources are released.

Each token assigned to the AMOUNT OF PROCESSOR AVAILABLE Place represents the available processing capacity in the infrastructure instantiated in the virtual machine of the private cloud. Each marking assigned to the TOTAL PROCESSOR Place represents the cloud processing infrastructure utilized. Each marking assigned to the amount of AMOUNT OF AVAILABLE MEMORY place represents the available memory capacity in the infrastructure instantiated in the virtual machine of the private cloud. Each token assigned to the TOTAL MEMORY Place represents the cloud memory infrastructure.

This performance model evaluates the impact of different types and levels of user requests on the big data environment set up in the private cloud by calculating processor utilization and memory utilization metrics. This assessment provides a means for planning infrastructures that meet certain workload levels with the desired quality of service.

The performance model calculates private cloud infrastructures' processor and memory utilization metrics (Table 1). Memory utilization (UM) represents the ratio between the memory used to service user requests and the total memory allocated to the data node. Processor utilization (UP) represents the fraction of time

that the processor remains busy serving user requests. This metric represents a percentage of the data node's total processing infrastructure utilization.

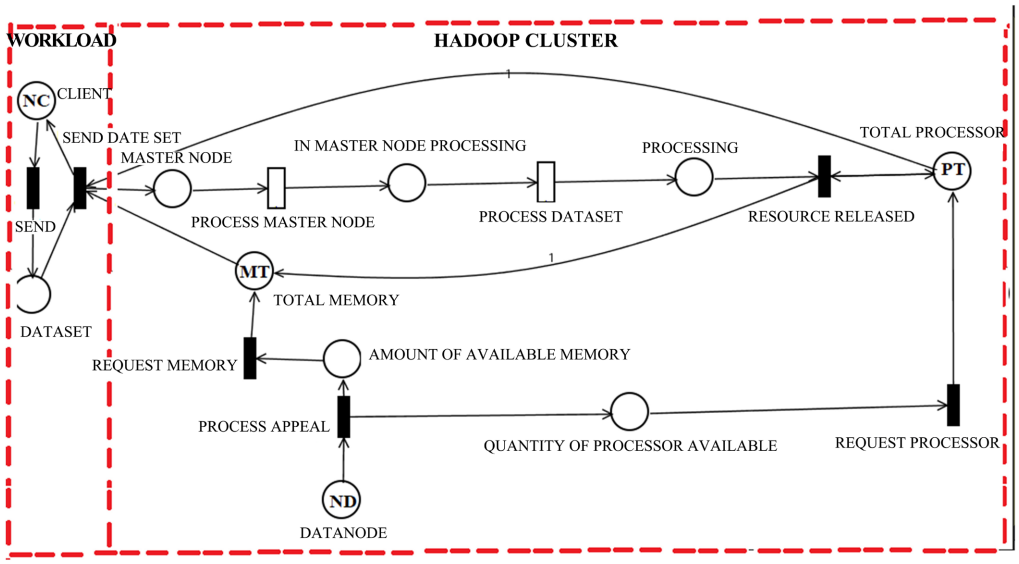

Figure 3. Performance model.

Table 1. Performance metrics.

Metric	Expression
Processador	((E{#MASTER NODE}) + (E{#PROCESSING MASTERNODE}) + (E{#PROCESSING})) × 100/(PROCESSOR TOTAL)
Memory	((MEMORY TOTAL) − (E{#MEMORYTOTAL})) × 100/(MEMORY − TOTAL)

4.2. Availability Model

In this section, RBD models were used to represent cloud system availability, and a proposed modeling strategy combines low-level RBD models with a high-level RBD model. Low-level RBD models are adopted to represent cloud infrastructure components and calculate high-level model parameters. The high-level model represents the cloud infrastructure and calculates the availability of this environment.

1) Cloud Platform Model—The Cloud Platform Model represents the components of cloud platforms through reliability block diagrams and calculates the availability of this infrastructure through the availability of its components (Figure 4).

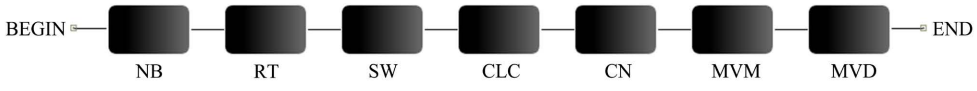

Figure 4. Cloud platform model.

The CloudStack platform is composed of the Cloud Controller (CLC), Node Controller (CN), virtual machines, and physical machines where the components of the CloudStack platform are configured and connected through no-break (NB), switch (SW), and a router (RT). The model parameters are the MTTF and MTTR of the CLC, CN, virtual machines, physical machines, no-break, switch, and router. The cloud platform model represents the components of these platforms through reliability block diagrams and calculates the cloud infrastructure's availability through its components' availability (Figure 4).

All cloud platform components must be operational for cloud computing to be operational. The operational mode of this cloud environment is $CLOUD\ PLATFORM\ MODEL\ OMCPM = (NB \wedge RT \wedge SW \wedge CLC \wedge CN \wedge MVM \wedge MVD)$, where NB, RT, SW, cloud controller, MVM, MVD are the no-break, router, switch, CLC, virtual machine of the master node, virtual machine of the data node. Figure 4 shows the RBD model adopted to estimate the availability of this platform's infrastructure.

2) Hardware Model—The Hardware RBD Model shown in (Figure 5) represents a computational system's processing and storage resources. The operational mode of this model is OMHARD = (MEM PROC DISC), where the MEM, PROC, and DISC are the memory, processor, and disk.

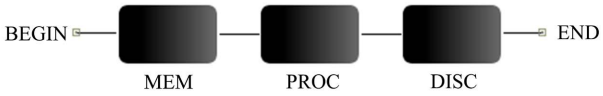

Figure 5. Hardware model.

3) Controller Model—The Controller Model (Figure 6) represents the controller of the private cloud. The operational mode of this model is $CONTROLLER\ MODEL\ OMCM = (S.O \wedge MG \wedge HARD)$, where the SO, MG, and HARD are the operating system, management module, and hardware components.

Figure 6. Controller model.

4) Node Controller Model—The Node Controller Model (Figure 7) represents the node controller of the private cloud. The operational mode of this model is $NODE\ CONTROLLER\ MODEL\ OMNCM = (HYPER\text{-}V \wedge S.O \wedge HARD)$, where HYPER-V, SO, HARD represents components Hyper-V, operating system, and hardware.

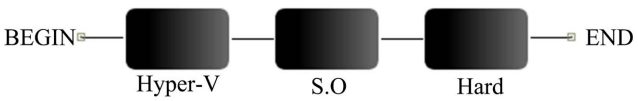

Figure 7. Node controller model.

5) VM Master Node Model—The VM Master Node Model (Figure 8) represents the virtual machine components of the master node. The operational mode of this model is *VM MASTER NODE MODEL OMVMMNM* = (*MN* ∧ *HYPER-V* ∧ *S.O*), where MN, HYPER-V, and SO represent the components of the master node, Hyper-V, the virtual machine's operating system.

Figure 8. VM master node model.

6) VM Data Node Model—The VM Data Node Model (Figure 9) represents the Data Node, Operating System, and HYPER-V components and the operational mode of this model is *VM DATA NODE MODEL OMVMDM* = (*DN* ∧ *S.O* ∧ *HYPER-V*), where the DN, SO, and HYPER-V represent the data node, operating system, and Hyper-V components of the data node.

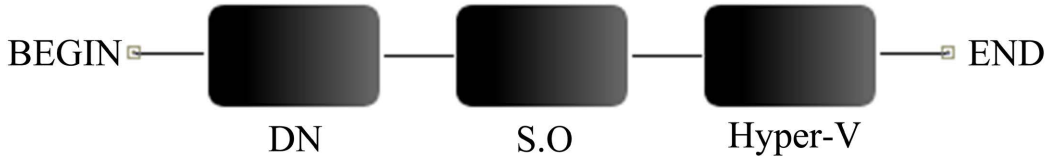

Figure 9. VM data node model.

5. Case Study

This section presents two case studies. Case Study 1 evaluates the performance of the Hadoop cluster configured in private cloud infrastructure, and Case Study 2 shows the availability of this environment using RBD models combined through a hierarchical modeling strategy.

5.1. Case Study 1

Case study 1 aims to evaluate the performance of the Hadoop cluster configured in the private cloud infrastructure according to the proposed methodology and models.

5.1.1. Understanding and Configuring the Big Data Environment in the Private Cloud

This section presents the environment adopted for evaluating the performance of big data environments in the cloud. The private cloud environment comprises 7 computers, an operating system without a graphical interface to reduce resource consumption, and the CloudStack [12] platform. These machines have the configuration shown in Table 2 considering the minimum configuration for installing the Hadoop cluster and cloud stack.

Table 2. Configuration of the machines that compose the private cloud.

Devices	Configuration
Memory	8 GB - 4 GB
Processor	I5 - I3
HD	1 TB
Hypervisor	KVM
S.O	CentoS 7
Cloud Platform	CloudStack

5.1.2. Planning Experiments in the Big Data Environment in the Private Cloud

In the design of experiments, factors with different levels were adopted. The service offering defines the capacity of the private cloud for the data nodes configured on the virtual machines. Table 3 shows the service offerings adopted for the design of experiments.

Table 3. Private cloud infrastructure service offering.

Service	Setup Offering
Small	Mem: 4 GB - Proc: 6 Cores - Storage: 1 TB
Medium	Mem: 6 GB - Proc: 6 Cores - Storage: 1 TB
Large	Mem: 8 GB - Proc: 8 Cores - Storage: 1 TB

The levels selected for the experimental planning are shown in Table 4. They were adopted according to the processing and memory capacity of the private cloud.

Table 4. Experimental planning scenario.

Scenario	Service Offering	Workload	Number of Data Nodes
1	Small	3 GB	3
2	Medium	4 GB	4
3	Large	5 GB	5

5.1.3. Workload Generation

To generate the workload, an emotional analysis of social network users on Twitter who made posts using words adopted by people with symptoms of depression was carried out. Between December 05 and 19, 2022, 5 GB of data were collected and converted into a dataset. The tool used in this data collection process was RStudio [15], which is an application that can capture data from Twitter with the execution of a script through a package called Twitterer [15]. According to the Pan American Health Organization [23], mental disorders are responsible for approximately 13% of the most common diseases in the world; more than 300

million people of all ages are affected by these disorders. The capture of publications on this social network was carried out according to emotional analysis parameters researched in the literature and on the website of the Ministry of Health of the federal government of Brazil. Words adopted by people who may have symptoms of depression are sad, anxious, depressive, depressed, schizophrenia, mental health, anxiety, therapy, mental illness, nervous, troubled, stressed, embarrassed, regretful, frustrated, and dissatisfied [24]. **Figure 10** shows the script used for data collection using the proposed tool. To generate the workload, an emotional analysis of social network users on Twitter who made posts using words adopted by people with symptoms of depression was carried out. Between December 05 and 19, 2022, 5 GB of data were collected and converted into a dataset. The tool used in this data collection process was RStudio [15], which is an application that can capture data from Twitter with the execution of a script through a package called Twitterer [15]. According to the Pan American Health Organization [23], mental disorders are responsible for approximately 13% of the most common diseases in the world; more than 300 million people of all ages are affected by these disorders. The capture of publications on this social network was carried out according to emotional analysis parameters researched in the literature and on the website of the Ministry of Health of the federal government of Brazil. Words adopted by people who may have symptoms of depression are sad, anxious, depressive, depressed, schizophrenia, mental health, anxiety, therapy, mental illness, nervous, troubled, stressed, embarrassed, regret, frustrated, and dissatisfied [24].

Figure 10 shows the script that was used for data collection using the proposed tool Rstudio, where lines 1, 2, 3, 4, 5, 6, and 7 of **Figure 10** provided the installation and request of the package required for access to social network data. Lines 9 to 13 indicate the supply of credentials (API KEY, API SECRET, ACCESS TOKEN, and ACCESS SECRET) for accessing social network data and authentication. Line 14 indicates the dataset collection period. Line 15 indicates which terms should be part of comments on social networks, which are hashtags related to emotional analysis. Lines 16 to 19 store in the variable tweets a search of 1000 posts from the social network. Line 20 has the function of structuring the social network data for processing in the Hadoop cluster [14] environment. Finally, line 23 saves the dataset to a file on HDFS, whereas lines 1, 2, 3, 4, 5, 6, and 7 of **Figure 10** provide the installation and request of the package required for access to social network data. Lines 9 to 13 indicate the supply of credentials (API KEY, API SECRET, ACCESS TOKEN, and ACCESS SECRET) for accessing social network data and authentication. Line 14 indicates the dataset collection period. Line 15 indicates which terms should be part of comments on social networks, which are hashtags related to emotional analysis. Lines 16 to 19 store in the variable tweets a search of 1000 posts from the social network. Line 20 has the function of structuring the social network data for processing in the Hadoop cluster [14] environment. Finally, line 23 saves the dataset to a file on HDFS.

5.1.4. Performance Measurement of Big Data Environments in the 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 measured using the Sysstat tool. Table 5 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 (sec) 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.

5.1.5. Statistical Analysis of Performance Metrics of Big Data Environments in the Private Cloud

In each experiment, the processor utilization and memory utilization metrics were collected according to the activity related to the methodology. Subsequently, the outliers in the performance metrics were removed using the Minitab tool. The averages of these metrics were calculated. The statistical analysis of the experiments was performed using Minitab software [18].

```
1   install.packges("twitteR")
2   library(twitteR)
3   require(twitteR)
4   Install.packages("ROAuth")
5   library("ROAuth")
6   install.packages(ROAuth")
7   require(RCurl)
8   # coloque suas chaves
9   api_key        <-"Informe a chave"
10  api_secret     <-"Informe a chave"
11  access_token   <-"Informe a chave"
12  access_secret <-"Informe a chave" setup_twitter_ouath(api_key,
access_token,access_secret)
13  interval<-dmy("19-12-2021")--dmy ("31-12-2021")
14  terms<-c("Stressed")
15  tweets<-searchTwitter(searchString = "#Stressed exclude:retweets",n = 1000)
16  tweetsTexts<-unlist(lapply(tweets, function(t)tStext))
17 tweetsTexts<-str_replace_all (tweetsTexts, "[^[:graph:]]","")
18  words<-unlist(strsplit(tweetsTexts,""))
19  words<-tolower(words)
20  tweets_df<-twListToDF(tweets)
21  getwd()
22  setwd("Local do arquivo")
```

Figure 10. Script used for data collection with Twitter.

Table 5. Processor and memory utilization metrics.

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

5.1.6. Mapping Performance Metrics of Big Data Environments in the Private Cloud

The performance metrics represented in the mapping are the processor and memory utilization of the data nodes through runtime configured in virtual machines of the private cloud [14].

5.1.7. 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 time 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 5.

5.1.8. Validation of the Performance Model of Big Data Environments in the Private Cloud

The proposed performance model was validated through a paired T-test that compares the mean difference between two independent samples; in this case, the metrics of processor utilization and memory utilization were measured in the 45 experiments and calculated through the proposed performance model [25]. 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, no statistical evidence exists to reject the equivalence hypothesis between measured values and those obtained from the performance model.

5.1.9. Analysis of New Scenarios of Big Data Environments in the Private Cloud

The new evaluated scenario considers the maximum workload supported by Hadoop cluster [14] with 3, 4, and 5 data nodes configured in the private cloud infrastructure with the large service offering, according to Table 6. This table presents the factors and levels of this study. These new scenarios were simulated using the proposed performance model. These new scenarios were simulated using the proposed performance model. Thus, the performance model's Workload subnet represented the workloads 10 GB, 15 GB, and 25 GB through the N parameter of the place Client.

Table 6. Private cloud infrastructure service offering.

Solutions	Service Offering	Workload	Number of Data Nodes
1	Large	10 GB	3, 4, 5
2	Large	15 GB	3, 4, 5
3	Large	25 GB	3, 4, 5

Table 7 presents the processor utilization and memory utilization of the data nodes configured in the private cloud virtual machines. It can be seen that greater workload intensities are applied, considering the big data environment with 10, 15, and 25 data nodes. Again, in this table, S means the scenarios, UPMod (%) represents the processor utilization obtained through the model, UMMod (%) is the average memory utilization obtained through the model, SF is the Service Offering, W means the Workload, and DN is the number of Data Nodes.

Table 7. Processor and memory utilization of new scenarios.

S	DN	SF	W	UPMod	UMMod
1	3	Large	10 GB	72.68	99.11
2	4	Large	10 GB	72.98	99.36
3	5	Large	10 GB	77.43	99.55
4	3	Large	15 GB	72.62	99.03
5	4	Large	15 GB	76.03	99.43
6	5	Large	15 GB	77.40	99.51
7	3	Large	25 GB	72.72	99.16
8	4	Large	25 GB	75.97	99.35
9	5	Large	25 GB	72.42	99.11

The workload impacts the variation of the processor and memory utilization metrics in Scenarios 1 to 9. In these Scenarios, the processor utilization metric did not reach saturation. However, the memory utilization metric reached saturation in all scenarios, with utilization values greater than 99 %, indicating a need to resize this resource to avoid performance loss in dataset is described in Table 7.

5.2. Case Study 2

Case Study 2 aims to evaluate the availability of the Hadoop cluster configured in the private cloud infrastructure. The next sections will present all the activities required for this evaluation.

5.2.1. Understanding and Configuring the Big Data Environment in the Private Cloud

The CloudStack platform was configured with 7 servers, one for the cloud controller (CLC) and the other for the node controller (CN). The servers that run the NC's services were configured with different virtual machines.

5.2.2. Generation of Availability Models

In this section, the Cloud Platform model was generated to evaluate the availability of big data environments in the private cloud by combining Hardware, Controller, Node Controller, VM Master Node, and VM Data Node models based on RBD according to the proposed hierarchical modeling strategy.

5.2.3. Parameterization of Availability Models

In this section, the RBD models were parameterized using the MTTFs and MTTR values of the private cloud and big data environment components. Figure 4 describes the Hadoop Cluster availability model configured in the private cloud. The Hardware model (Figure 5) represents the memory, processor, and disk and is adopted to calculate the MTTF and MTTR of the computing system, as per Table 8 [26]. The MTTF and MTTR values calculated for the computational system are 279,069 hours and 8 hours, respectively.

Table 8. MTTF and MTTR values of the computational system components.

Components	MTTF	MTTR
Processor	1,500,000	8
Memory	480,000	8
Disk	1,200,000	8

The Controller Model (**Figure 6**) is composed of the cloud controller components and is adopted to calculate the MTTF and MTTR of the cloud controller, considering **Table 9**. The cloud controller's calculated MTTF and MTTR values are 19,530 hours and 8 hours [26].

Table 9. MTTF and MTTR values of the cloud controller components.

Components	MTTF	MTTR
Operating System	42,000	8
Management Module	42,000	8
Computational System	279,069	8

The Node Controller Model (**Figure 7**) is composed of the virtual machine monitor, operating system, and computational system and is adopted to calculate the MTTF and MTTR of the node controller, as shown in **Table 10**. The MTTF and MTTR values calculated for the node controller are 2763 and 8 hours [26].

Table 10. MTTF and MTTR values of the node controller components.

Components	MTTF	MTTR
Operating System	42,000	8
Virtual Machine Monitor	2990	8
Computational System	279,069	8

The VM Master Node Model (**Figure 8**) is composed of the master node components, virtual machine monitor, and operating system and is adopted to calculate the MTTF and MTTR of the master node virtual machine, as shown in **Table 11**. The MTTF and MTTR values calculated for the master node virtual machine are 2,617 hours and 8 hours [26].

The VM Data Node Model (**Figure 9**) is composed of the virtual machine monitor, operating system, and computing system and is adopted to calculate the MTTF and MTTR of the data node virtual machine, as shown in **Table 12**. The values of MTTF and MTTR calculated for the data node virtual machine are 2,767 hours and 8 hours [26].

Table 11. MTTF and MTTR values of the VM master node components.

Components	MTTF	MTTR
Operating System	42,000	8
Virtual Machine Monitor	2990	8
Master Node Components	42,000	8

Table 12. MTTF and MTTR values of the VM DATA node components.

Components	MTTF	MTTR
Operating System	42,000	8
Virtual Machine Monitor	2990	8
Data Node Components	329,067	8

The Cloud Platform Model (Figure 4) comprises the no-break, router, switch, controller, node controller, master node virtual machine, and data node virtual machine. The cloud platform availability is calculated according to Tables 8-12 [26]. The cloud platform availability is 98.79%.

5.2.4. Analysis of New Scenarios of Big Data Environments in the Private Cloud

In this section, we can evaluate the impact on the availability of the percentage variation of 50% for more and 50% for less of the MTTF of each private cloud component, according to Table 13. The cloud platform availability is 98.79%, but when there is a variation of more than 50% of the MTTF of each component, availability increases to 99.19%. The availability value is reduced to 98.00% when there is a variation of minus 50% in each component's MTTF value.

Table 13. MTTF and MTTR Values with Percentage Variation of 50% for more and 50% for Less of the MTTF of each Component of the Private Cloud Platform.

Components	MTTF	MTTR	+50% MTTF	−50% MTTF
NB	329067.64	8	493601.46	164533.5
RT	42000.0	8	63000.0	21000.0
SW	2990.0	8	4485.0	1495.0
CLC	19,530	8	29,295	9765
CN	2763	8	4144.5	1381.5
MVM	2617	8	3925.5	1308.5
MVD	2767	8	4150.5	1383.5

6. Discussion

Previous works [27] [28] have proposed stochastic models based on Petri nets for performance evaluation of cloud environments. Similarly, some works [29] [30] have presented RBD-based models for the dependability evaluation of cloud

environments. However, this work presents performance and availability models for evaluating big data environments in cloud computing. In addition, this work also presents methodologies for performance evaluation and availability evaluation of big data environments in cloud computing, considering different big data environments configured in private clouds with various configurations.

Furthermore, the case studies were developed based on the proposed methodologies. These case studies focus on the performance evaluation and availability assessment of a private cloud environment configured with the OpenStack platform and a big data environment configured with the Hadoop cluster. However, different scenarios can be evaluated based on these methodologies.

The proposed work has limitations related to the performance model since metrics such as response time and throughput are not evaluated. Availability models can be expanded to assess the impact of maintenance activities on the availability of the big data environment in cloud computing.

7. Conclusions

The contributions of this work were the proposal of a methodology and a stochastic model for the performance evaluation of big data environments in the private cloud. This work provides a 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. To generate the workload, sentiment analysis was performed on Twitter users' posts with words that indicated symptoms of depression.

The proposed performance model was based on stochastic Petri nets, and the processor utilization and memory utilization of the Hadoop cluster in the private cloud were evaluated, considering different service offerings, workloads, and the number of data nodes. Case study 1, 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 processor and memory utilization metrics obtained through the validated performance model showed that the memory resource saturated when the large service offering was adopted for setting 3, 4, or 5 data nodes, with 10 GB, 15 GB, and 25 GB workloads. These results demonstrated that workload is the biggest 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.

This work also presents a hierarchical modeling strategy and RBD models for evaluating the availability of big data environments in the private cloud. The computational system model, controller model, controller node model, virtual machine of the master node model, and virtual machine of the data node model are adopted to calculate the parameters of the cloud computing model, and this model

calculates the availability of the big data environment in the private cloud. In Case Study 2, this modeling strategy was adopted to evaluate the impact of varying the MTTF value on the availability of the big data environment in cloud computing.

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

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Marinescu, D.C. (2017) Cloud Computing: Theory and Practice. Morgan Kaufmann.

[2] Bertoncello, G. (2018) Um Estudo Sobre a Performance de Aplica¸c˜oes Big Data com Deep Learning. Universidade Federal do Rio Grande do Sul.

[3] Veiga, J., Exposito, R.R., Pardo, X.C., Taboada, G.L. and Tourifio, J. (2016) Performance Evaluation of Big Data Frameworks for Large-Scale Data Analytics. 2016 *IEEE International Conference on Big Data* (*Big Data*), Washington, 5-8 December 2016, 424-431. https://doi.org/10.1109/bigdata.2016.7840633

[4] Melo, C., Matos, R., Dantas, J. and Maciel, P. (2017) Capacity-Oriented Availability Model for Resources Estimation on Private Cloud Infrastructure. 2017 *IEEE* 22*nd Pacific Rim International Symposium on Dependable Computing* (*PRDC*), Christchurch, 22-25 January 2017, 255-260. https://doi.org/10.1109/prdc.2017.49

[5] Dantas, J.R. (2018) Planejamento de infraestrutura de nuvens computacionais para serviço de *VoD streaming* considerando desempenho, disponibilidade e custo. Ph.D. Thesis, Universidade Federal de Pernambuco.

[6] Oliveira, A.S. (2017) SIMF: Um Framework de Inje¸c˜ao e Monitoramento de Falhas de Nuvens Computacionais Utilizando SPN. Universidade Federal de Pernambuco.

[7] Jain, R. (1991) The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling. Wiley.

[8] Lilja, D.J. (2005) Measuring Computer Performance: A Practitioner's Guide. Cambridge University Press.

[9] Menasce, D.A., Almeida, V.A. and Dowdy, L.W. (2004) Performance by Design: Computer Capacity Planning by Example. Prentice Hall Professional.

[10] Rahman, H., Shahina, B. and Ahmed, M.U. (2016) Ins and Outs of Big Data: A Review. *The* 3*rd EAI International Conference on IoT Technologies for HealthCare*, Västerås, 18-19 October 2016, 44-51. https://doi.org/10.1007/978-3-319-51234-1_7

[11] Kuo, W. and Zuo, M. (2003) Optimal Reliability Modeling—Principles and Applications. Wiley.

[12] Cloudstack. https://cloudstack.apache.org/

[13] Openstack. https://www.openstack.org

[14] Apache Hadoop (2024) 10 Charts That Will Change Your Perspective of Big Data's Growth. https://hadoop.apache.org/

[15] Rstudio. https://posit.co/downloads/

[16] CentOS. https://www.centos.org/

[17] Kuo, W. and Zuo, M.J. (2003) Optimal Reliability Modeling: Principles and Applications. Wiley.

[18] Minitab. https://www.minitab.com/pt-br/

[19] Khalifa, A. and Eltoweissy, M. (2013) Collaborative Autonomic Resource Management System for Mobile Cloud Computing. *Proceedings of the Fourth International Conference on Cloud Computing GRIDs, and Virtualization*, Valencia, 27 May-1 June 2013, 115-121.

[20] Yee, S.-T. and Ventura, J.A. (2000) Phase-Type Approximation of Stochastic Petri Nets for Analysis of Manufacturing Systems. *IEEE Transactions on Robotics and Automation*, **16**, 318-322. https://doi.org/10.1109/70.850650

[21] Xie, M., Dai, Y.S. and Poh, K.L. (2004) Computing System Reliability: Models and Analysis. Kluwer Academic Plenum Publishers.

[22] Eucalyptus (2024) Amazon Web Services. Eucalyptus Open Source Cloud Computing Infrastructure an Overview. https://aws.amazon.com

[23] Opas. https://www.paho.org/pt/brasil/

[24] Saude Gov. https://www.gov.br/saude/pt-br

[25] Gupta, B.C. and Guttman, I. (2014) Statistics and Probability with Applications for Engineers and Scientists. Wiley.

[26] Intel (2024) Processadores Intel® Core™ i5. https://www.intel.com.br/content/www/br/pt/products/details/processors/core/i5/docs.html?s=Newest/

[27] Yadav, R.R., Campos, G.A.S., Sousa, E.T.G. and Lins, F.A. (2019) A Strategy for Performance Evaluation and Modeling of Cloud Computing Services. *Revista de Informática Teórica e Aplicada*, **26**, 78-90. https://doi.org/10.22456/2175-2745.87511

[28] Ali, M.R., Ahmad, F., Chaudary, M.H., Khan, Z.A., Alqahtani, M.A., Alqurni, J.S., *et al.* (2021) Petri Net Based Modeling and Analysis for Improved Resource Utilization in Cloud Computing. *PeerJ Computer Science*, **7**, e351. https://doi.org/10.7717/peerj-cs.351

[29] De Sousa, E.T.G. and Lins, F.A.A. (2018) Modeling Strategies to Improve the Dependability of Cloud Infrastructures. In: Márquez, F.P.G. and Papaelias, M., Eds., *Dependability Engineering*, InTech, Vol. 7. https://doi.org/10.5772/intechopen.71498

[30] Dantas, J., Matos, R., Araujo, J., Oliveira, D., Oliveira, A. and Maciel, P. (2016) Hierarchical Model and Sensitivity Analysis for a Cloud-Based VoD Streaming Service. 2016 46*th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop* (*DSN-W*), Toulouse, 28 June-1 July 2016, 10-16. https://doi.org/10.1109/dsn-w.2016.23