UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEPARTAMENTO DE COMPUTAÇÃO
PROGRAMA DE PÓS–GRADUAÇÃO EM INFORMÁTICA APLICADA

TARCÍSIO JOSÉ ROLIM FILHO

AVALIAÇÃO DE DESEMPENHO, DISPONIBILIDADE E PERFORMABILIDADE DE AMBIENTES BIG DATA NA NUVEM PRIVADA

RECIFE – PE

2022

TARCÍSIO JOSÉ ROLIM FILHO

AVALIAÇÃO DE DESEMPENHO, DISPONIBILIDADE E PERFORMABILIDADE DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática - DEINFO - Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre.

ORIENTADOR: PROF[a]. ERICA TEIXEIRA GOMES DE SOUZA

RECIFE – PE

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

143a Filho, Tarcísio José Rolim
AVALIAÇÃO DE DESEMPENHO, DISPONIBILIDADE E PERFORMABILIDADE DE AMBIENTES BIG DATA NA NUVEM PRIVADA / Tarcísio José Rolim Filho. - 2023.
87 f.

Orientadora: ERICA TEIXEIRA GOMES DE SOUZA.
Inclui referências.

Dissertação (Mestrado) - Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, 2023.

1. Cloud Computing. 2. Big data. 3. Hadoop. 4. Performance Evaluation Dependability. 5. Evaluation Stochastic Petri Nets. I. SOUZA, ERICA TEIXEIRA GOMES DE, orient. II. Título

CDD 004

TARCÍSIO JOSÉ ROLIM FILHO

AVALIAÇÃO DE DESEMPENHO, DISPONIBILIDADE E PERFORMABILIDADE DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática - DEINFO - Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre.

Aprovada em:

BANCA EXAMINADORA

PROFª. ERICA TEIXEIRA GOMES DE SOUZA (Orientador)
UFRPE
Departamento de Computação

PROFº. FERNANDO ANTÔNIO AIRES LINS
UFRPE
Departamento de Computação

ROBSON WAGNER ALBUQUERQUE DE MEDEIROS
UFRPE
Departamento de Computação

À minha família.

À professora Erica Sousa pela orientação.

À Maria Cecília Monteiro Rolim.

Agradecimentos

Agradeço a DEUS, por sempre estar presente na minha vida em todos os momentos. Agradeço a minha esposa, Ana Carolina, pelas palavras de incentivo, compreensão e carinho.
A minha família por todo incentivo.

A minha orientadora, Profª Dra. Érica Sousa, pela orientação, dedicação, paciência, incentivo, comentários que foram essenciais para esta dissertação. Também gostaria de agradecer por todas as oportunidades de crescimento profissional e pessoal proporcionadas desde o início do mestrado.

Agradeço à FACEPE pelo suporte financeiro para o desenvolvimento deste trabalho.

Agradeço ao Departamento de Computação da Universidade Federal Rural de Pernambuco, pelo apoio e suporte durante a elaboração deste trabalho.

"Por vezes sentimos que aquilo que fazemos
não é senão uma gota de água no mar. Mas
o mar seria menor se lhe faltasse uma gota".
(Madre Teresa de Calcuta)

Resumo

A computação em nuvem permite escalabilidade a um custo menor para análise de dados em um ambiente de big data. Esse paradigma considera o dimensionamento de recursos para processar diferentes volumes de dados minimizando o tempo de resposta da manipulação de *big data*. Este trabalho propõe a avaliação de desempenho, disponibilidade e performabilidade de ambientes de *big data* na nuvem privada por meio de uma metodologia e modelos estocásticos e combinatoriais considerando métricas de desempenho, como tempos de execução, utilização de processador, memória e disponibilidade do ambiente big data na nuvem privada. A metodologia proposta considera atividades objetivas e modelagem de desempenho e disponibilidade para avaliar o ambiente na nuvem privada. O modelo de desempenho baseado em redes de Petri estocásticas é proposto para avaliar a utilização de processadores e memória do ambiente big data da nuvem privada, cujo representa o envio de conjuntos de dados para o *cluster Hadoop* com diferentes configurações. Diagramas de bloco de confiabilidade são adotados para avaliação de disponibilidade de ambientes *big data* na nuvem privada. Três estudos de caso baseados na plataforma CloudStack e *cluster Hadoop* são adotados para demonstrar a viabilidade da metodologia e modelo propostos. Para gerar a carga de trabalho, foi feita uma análise dos sentimentos dos usuários do *Twitter* que fizeram postagens que trazem indícios de sintomas de depressão. O Estudo de caso 1 proporcionou a avaliação das métricas de desempenho dos cenários, na utilização do processador, memória e tempo de resposta do Hadoop cluster na nuvem privada, considerando diferentes ofertas de serviço, carga de trabalho e número de *data nodes*. O Estudo de caso 2 proporcionou a avaliação da disponibilidade através de métricas de ambientes big data na nuvem privada. O estudo de caso 3 apresenta os resultados de performabilidade através do método para avaliação de performabilidade de ambientes *big data* na nuvem privada.

Palavras-chave: Computação em Nuvem. *big data*. *Hadoop*. Avaliação de Desempenho. Confiabilidade. Redes de Petri Estocástica.

Abstract

Cloud computing enables scalability at a lower cost for analytics data in a big data environment. This paradigm considers the dimensioning of resources to process different volumes of data minimizing the response time of the big data manipulation. This performance evaluation proposal work, availability and performability of big data environments in the private cloud through a methodology and stochastic and combinatorial models considering performance measurements, such as execution, processor utilization, memory and availability of the big data environment in the private cloud. The proposed methodology considers objective activities and modeling performance and availability metrics to evaluate the Hadoop cluster in the private cloud. The performance model based on stochastic Petri nets proposed to evaluate the private memory usage representative of the cloud big data environment sending datasets to the Hadoop cluster with different configurations. Diagrams of loyalty block are accepted for evaluating the availability of big marriages data in the private cloud. Three case studies on the CloudStack platform and cluster Hadoop are adopted to demonstrate the methodology of the proposed methodology and model. To generate the workload, an analysis of the feelings of the users of the Twitter that postulations with words that emerge from depression. The evaluation of the performance measurements of the studies carried out, on processor utilization, memory and response time of the Hadoop cluster in the cloud private, considering different service offerings, workload and number of data we. Assessment Study 2 provided an assessment of availability through case measurements of big data environments in the private cloud. Case study 3 presents the performability results through the method for performance evaluation of big data environments in the private cloud.

Keywords: Cloud Computing, Big data, Hadoop, Performance Evaluation Dependability, Evaluation Stochastic Petri Nets.,

Lista de Figuras

Lista de tabelas

Lista de abreviaturas e siglas

IOT	*Internet-Of-Things*
TI	Tecnologia da Informação
SPN	*Stochastic Petri Net*
RBD	*Reliability Block Diagram*
PNs	*Petri Nets*
SAS	*Software-as-a-Service*
IAAS	*Infrastructure-as-a-service*
PAS	*Platform-as-a-service*
ERP	*Enterprise Resource Planning*
CRM	*Customer relationship management*
MTTR	*Mean Time to Repair*
MTTF	*Mean Time to Failure*
CPU	*central processing unit*
DL	*Deep Learning*
SRN	redes de Petri de Recompensa
CTMC	Cadeia de Markov de tempo contínuo
AD	Avaliação de dependabilidade
AVD	avaliação de desempenho
AP	Avaliação de performabilidade
KVM	*Kernel-based Virtual Machine*
BD	*Big data*

NP	Nuvem privada
VM	Máquina virtual
NC	Número de Cliente
TP	Tempo de envio
TT	Transição temporizada
ND	Número de datanode
MT	Memória Total
PT	Processador Total
UM	Utilização da memória
UP	Utilização do processador
MEM	Memória
PPROC	Processador
DISC	Disco
S.O	Sistema Operacional
MG	Módulo de Gerenciamento
HARD	*Hardware*
NB	*Nobreak*
RT	Roteador
SW	*Switch*
CLC	Controlador de nuvem
CN	Controlador de nó
MVM	Máquinan Virtual do Master Node
MVD	Máquinan Virtual do Data Node

OM	Módulo Operacional
DN	*Datanode*
MN	*MasterNode*
HD	*Hard Disk*
GB	*Giga Byte*
OPAS	Organização Pan-americana de Saúde
OPAS	*Hadoop Distributed File System*
OF	Oferta de serviço
CG	Carga de trabalho
EC	Estudo de caso

Sumário

1 INTRODUÇÃO

A computação em nuvem é uma tecnologia que permite a distribuição dos seus serviços de computação e o acesso online a eles sem a necessidade de instalar programas. Justamente por não necessitar da instalação de programas, ou do armazenamento de dados, o conceito originado do inglês *cloud computing* faz alusão à computação em nuvem (OUTAMAZIRT et al., 2018). Com isso, seus serviços podem ser acessados de maneira remota, de qualquer lugar do mundo e a hora que for desejado. A distribuição dos serviços é feita por meio de uma plataforma de serviços *cloud* via Internet com uma definição de preço conforme o uso.

Portanto, a computação em nuvem pode proporcionar inovações mais rápidas, recursos flexíveis e economia em escala. Por não precisar de uma máquina potente, pois esse serviço oferece acesso rápido a recursos de TI flexíveis, ação em escala global, aumento de produtividade, melhor desempenho e maior segurança(MARINESCU, 2017).

Vários paradigmas de computação, como *Edge* que é computação de borda em que uma mudança de perspectiva em relação a Cloud Computing, uma vez que nesse tipo de solução todo o processamento de dados acontece na borda, isto é, nos próprios dispositivos utilizados pelos usuários e *Fog computing* que significa computação em neblina, surgiram para oferecer suporte a serviços sensíveis a atrasos e sensíveis ao contexto. Ao combinar dispositivos de borda, servidores de névoa e computação em nuvem, as empresas podem construir uma infraestrutura de IoT hierárquica, usando a arquitetura orquestrada *Edge–Fog–Cloud* para melhorar o desempenho, a confiabilidade e a disponibilidade dos ambientes de IoT, basicamente, a ideia trata justamente da mistura entre soluções na borda e na nuvem. A essência da Fog é promover uma arquitetura descentralizada. As aplicações e o gerenciamento são distribuídos de maneira inteligente entre a fonte dos dados e a nuvem. (SCHENFELD, 2017).

A computação em nuvem proporciona vários benefícios como a redução de custo com a infraestrutura de TI para armazenamento de grandes conjuntos de dados, escalabilidade com rápida expansão da infraestrutura e recuperação de desastres. (TAMURA; YAMADA, 2015).

A importância de pesquisar a avaliação de desempenho, disponibilidade e

performabilidade é que na avaliação de desempenho permite a avaliação do impacto da variação da carga de trabalho submetida à nuvem computacional com diferentes configurações de software e hardware. Esse modelo de desempenho possibilita o cálculo do tempo de resposta e da utilização de recursos das infraestruturasde nuvem computacional. Nos modelos de disponibilidade dos componentes da nuvem computacional as métricas de desempenho e de dependabilidade podem ser combinadas para seleção de infraestruturas de nuvem que atendam aos requisitos dos clientes.

Por outro lado, o big data surgiu para atender a demanda por novas técnicas que permitam o processamento de informações com alto desempenho e disponibilidade. As ferramentas de *Big Data* tornam a coleta, processamento e visualização de dados mais simples, padronizadas e eficazes (MACHADO, 2018a). Os benefícios da computação em nuvem para o ambiente big data é a capacidade de fornecer uma solução escalável e adaptável para grandes conjuntos de dados e análise de negócios. A nuvem privada é um modelo de computação em nuvem em que a infraestrutura é dedicada a uma única organização do usuário. Por isso, a quantidade de máquinas virtuais na nuvem privada e o tamanho da carga de trabalho, podem ser propostas para um melhor gerenciamento da infraestrutura de nuvem privada.

1.1 MOTIVAÇÃO E JUSTIFICATIVA

A computação em nuvem é um conjunto de recursos virtualizados facilmente utilizáveis e acessíveis, tais como hardware, software, plataformas de desenvolvimento e serviços. Estes recursos podem ser dinamicamente reconfigurados para se ajustarem à uma carga de trabalho variável, permitindo a otimização do seu uso. A implementação de uma nuvem privada no âmbito de uma empresa pode agregar diversos benefícios tais como: melhorias no aproveitamento dos recursos, redução dos custos com manutenção, redução do consumo de energia e permitir maior controle das configurações da nuvem.(WANG Y.; KUNG, 2018).

De acordo com o site Statista (STATISTA, 2022), o mercado global de análise de big data crescerá com uma taxa de crescimento anual composta de quase 30% nos

próximos anos, com receita atingindo mais de 68 bilhões de dólares até 2025.

Nesses ambientes o planejamento da infraestrutura de nuvem privada é uma atividade crítica porque permite que o provedor de nuvem tenha recursos suficientes para alocar e liberar dinamicamente, quando sujeito a níveis variados de demandas do usuário. Esse planejamento também permite que a infraestrutura de nuvem privada seja dimensionada para suportar cargas de trabalho de alto nível com tempos de resposta aceitáveis. A avaliação de desempenho da nuvem privada permite atender aos requisitos do usuário em diferentes níveis, mantendo a qualidade do serviço oferecido

Com isso, o planejamento da infraestrutura de nuvem privada por ser difícil de mensurar e avaliar torna-se uma atividade essencial pois possibilita que uma empresa possa ter recursos suficientes para alocá-los e disponibilizar dinamicamente, quando submetido aos diferentes níveis de requisição dos clientes. Esse planejamento também permite o dimensionamento significativo da infraestrutura de nuvem privada para suportar altos níveis de carga de trabalho com tempos de resposta. Contudo, modelos de dependabilidade e performabilidade são aliados a modelos de tomada de decisão que auxiliam usuários na análise e escolha de infraestruturas (ARAÚJO, 2019).

Desta forma, o problema de pesquisa que motiva esta dissertação é deescrito na seguinte pergunta: "Como avaliar o desempenho, disponibilidade e performabilidade de ambientes big data na nuvem privada?".

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

Este trabalho tem o objetivo de propor uma estratégia baseada em uma metodologia para avaliação de desempenho, disponibilidade e performabilidade, modelos *Stochastic Petri Net* (SPN) e *Reliability Block Diagram* (RBD) para avaliar o desemepnho, a disponibilidade e performabilidade de ambientes big data na nuvem privada.

1.2.2 OBJETIVO ESPECÍFICO

Os objetivos específicos deste trabalho estão descritos a seguir:

- Propor uma metodologia para avaliação de desempenho, disponibilidade e

performabilidade em ambientes *big data* na nuvem privada;

• Construir um modelo de desempenho baseado em Rede de Petri Estocástica (SPN) para ambientes de big data na nuvem;

• Criar modelos de disponibilidade do ambiente big data na insfraestrutura utilizando formalismo de Diagrama de blocos de confiabilidade (RBD).

• Propor uma estratégia de modelagem hierárquica para avaliação de disponibilidade de ambientes *big data* na nuvem.

1.3 ESTRUTURA DO DOCUMENTO

Este trabalho está dividido em 7 capítulos, os quais serão brevemente destacados nesta seção.

O Capítulo 2 apresenta a fundamentação teórica do trabalho proposto. Esse capítulo apresenta uma visão geral sobre a computação em nuvem e *big data*, além de conceitos básicos sobre avaliação de desempenho, disponibilidade e performabilidade. Uma introdução redes de Petri, diagramas de bloco de confiabilidade e conceitos básicos sobre big data, avaliação de desempenho, avaliação de disponibilidade, avaliação de performabilidade, técnicas de aproximação de fases e planejamentos de experimentos é apresentada.

O Capítulo 3 apresenta os trabalhos relacionados à avaliação de desempenho, dependabilidade e performabilidade de computação em nuvem ou *big data*.

O Capítulo 4 apresenta a metodologia para avaliação de desempenho disponibilidade e performabilidade em ambientes *big data* em nuvens privadas.

O Capítulo 5 apresenta os modelos de desempenho e disponibilidade, além da estratégia de modelagem hierárquica proposta.

O Capítulo 6 apresenta três estudos de caso para a aplicação da metodologia, estratégia de modelagem e modelos propostos. Esses estudos de caso são baseados no planejamento de ambientes big data na nuvem privada são configurados.

O capítulo 7 apresenta as conclusões, limitações e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo introduz os conceitos básicos para compreensão do trabalho proposto, inicialmente apresenta uma introdução à computação em nuvem. Em seguida apresenta conceitos básicos sobre *big data*. Em seguida, apresenta os conceitos básicos sobre avaliação de desempenho, avaliação de dependabilidade e avaliação de performabilidade. Em sequência, apresenta os principais conceitos sobre redes de Petri (Petri Nets - PNs) e apresenta uma de suas extensões, as redes de Petri estocásticas (*Stochastic Petri Nets - SPNs*). Em continuidade, são introduzidos conceitos sobre a técnica de aproximação de fases. Finalmente, esse capítulo apresenta uma introdução sobre planejamento de experimentos.

2.1 COMPUTAÇÃO EM NUVEM

A computação em nuvem permite que os usuários foquem em suas próprias atividades, deixando a manutenção e o gerenciamento da infraestrutura do ambiente para provedores de nuvem realizar aplicativos e serviços pela internet. A computação em nuvem oferece serviços de armazenamento de dados sem a necessidade do cliente dispor de uma infraestrutura dedicada (KHALIFA; ELTOWEISSY, 2013). Portanto, tanto a infraestrutura quanto a interface da aplicação são disponibilizadas por um provedor de serviços. Estes podem ser adaptados às necessidades do cliente sem que exista a participação na instalação, configuração ou manutenção do produto.
Existem muitas classificações para as abordagens em nuvem, temos assim as que se referem ao modelo de serviços (SaaS, IaaS e PaaS) e ao modelo de implantação (pública, privada, híbida e comunitária). As soluções em nuvem, quanto ao seu modelo de serviços, são divididas em 3 camadas.

Infraestrutura como Serviço (IaaS) é comumente utilizada por gerentes de sistemas na criação de máquinas virtuais, sistemas operacionais e memórias virtuais etc. É a camada mais profunda da nuvem. Seu principal objetivo é prover um ambiente de armazenamento, processamento e utilização de memória sob demanda, de fácil compreensão ao usuário final e com múltiplos recursos disponíveis.

Plataforma como Serviço (PaaS) é a capacidade dada aos usuários finais de acessar

plataformas de desenvolvimento e implantação sob demanda fornecidas pelo provedor pela Internet.A Plataforma é mais usada pelos desenvolvedores na integração de aplicações, experimentações e inserção de *frameworks*.

Sua presença assegura o desenvolvimento de aplicações sem preocupação com a capacidade de servidores, realização de testes e análises de dados, integração com bancos de dados etc.

Software como Serviço (SaaS) é a camada mais externa e perceptível da nuvem. O *Software* como Serviço (SaaS) é composto por um grupo de aplicativos executados diretamente no ambiente virtual, por meio de uma interface disponibilizada na web. É mais utilizado pelo usuário final para acesso a e-mails ou aplicativos de escritório (ERP, CRM e plataforma de assinatura eletrônica, por exemplo).

Quanto ao modelo de implantação a nuvem podem ser privada, pública, comunitária e híbrida.

Uma nuvem privada consiste em recursos de computação em nuvem usados exclusivamente por uma única empresa ou organização. A nuvem privada pode estar localizada fisicamente no *datacenter* local da organização ou pode ser hospedada por um provedor de serviços. Mas em uma nuvem privada, os serviços e a infraestrutura são sempre mantidos na rede privada e o hardware e o *software* são dedicados unicamente à organização.

Em uma nuvem pública, os recursos de nuvem (como servidores e armazenamento) pertencem a um provedor de serviço de nuvem, e são operados por ele e entregues pela Internet. Podemos compartilhar os mesmos dispositivos de *hardware*, de armazenamento e de rede com outras organizações ou "locatários"da nuvem, acessar serviços e gerenciá-los pela conta usando um navegador da internet.

A Nuvem híbrida é um tipo de computação em nuvem que combina uma infraestrutura local ou nuvem privada com uma nuvem pública. As nuvens híbridas permitem que os dados e aplicativos se movam entre os dois ambientes. Muitas organizações adotam a abordagem de nuvem híbrida devido a exigências comerciais, por exemplo, para atender a requisitos regulatórios e de soberania de dados, aproveitar ao máximo o investimento em tecnologia local ou lidar com problemas envolvendo latência baixa.

A infra-estrutura de uma nuvem comunitária é compartilhada por várias organizações que partilham interesses como a missão, requisitos de segurança, políticas, entre outros. Pode ser administrada pelas próprias organizações ou por um terceiro e pode existir no ambiente da empresa ou fora dele. Essa opção pode oferecer um nível maior de privacidade, segurança e/ou conformidade com a política da organização e pode ser gerenciada pelas organizações envolvidas ou por um terceiro. Dessa forma, esse modelo pode ser economicamente mais atrativo, pois os recursos (armazenamento, estações de trabalho) utilizados e compartilhados na comunidade já representam um bom retorno em investimento.

2.2 BIG DATA

O *big data* surgiu para atender a demanda por novas técnicas que permitissem o processamento de informações com alto desempenho e disponibilidade (COLUMBUS, 2018). As ferramentas de *Big Data* tornam a coleta, processamento e visualização de dados mais simples, padronizadas e eficazes. As tomadas de decisão baseadas em análise de dados, estão cada vez mais presentes em ambientes de negócios e ambientes acadêmicos. Tecnologias de processamento e gerenciamento de dados, como o *Hadoop* (HADOOP, 2022) e MapReduce que é um mecanismo de análise de dados amplamente usado para Big Data(DEAN J.AND GHEMAWAT, 2010), permitem geração de informações a partir de um grande conjunto de dados, considerando critérios de qualidade, satisfação de *stakeholders* e utilização de recursos. (MACHADO, 2018b).

De acordo com Maheshwari (MAHESHWARI, 2018), operações de Big data demandam capacidade que excede os atuais sistemas de armazenamento convencionais e esses ambientes podem ser caracterizados por 5vs: Volume, velocidade, variedade, veracidade e valor.

De acordo com Rahman (RAHMAN et al., 2016), big data é caracterizado pelas três dimensões ou 3V chamado volume, velocidade, veracidade, variedade, validade e valor. De acordo com o estudo, três dimensões adicionais veracidade, validade e valor consideraram com "6 Vs" de *big data*. O 1º V é o Volume que diz respeito da quantidade

de dados gerados que está aumentando extremendamente a cada dia.
O segundo V é a velocidade que veio à tona devido a mais e mais dados e é fornecidos aos usuários imediatamente sempre que necessário para processamento em tempo real.
Variedade é o 3º V considerado devido ao enorme crescimento das fontes de dados que são necessários para análise. A veracidade que inclui a confiança nas informações recebidas é frequentemente citada como 4ª dimensão V, além de *big Data*. O 5º V é a validade não envolve apenas garantindo medições precisas, mas também a transparência de suposições e conexões por trás do processo. O 6º V é o valor que refere-se a grandes volumes recentes de dados medidos em *exabytes*, *petabytes* ou classificação mais alta de dados e altamente valioso para pesquisa institutos e indústrias.

Diante deste fato, diversas tecnologias estão sendo abordados e desenvolvidos, tais como *Hadoop* e *MapReduce* (ASSUNÇÃO et al., 2015), (HASHEM et al., 2015a), (DEAN; GHEMAWAT, 2010a), (OUSSOUS et al., 2018). O *Hadoop* é um *framework* que processa grandes quantidades de dados através de mecanismos distribuídos como o *MapReduce* que é um modelo de programação que permite o processamento de dados massivos em um algoritmo paralelo e distribuído, geralmente em um cluster de computadores (DEAN; GHEMAWAT, 2010b) e foi desenvolvido para contribuir com os desafios de ambientes *Big Data* em analisar dados gerados por diferentes fontes, como redes sociais, dados de saúde e dados climatológicos (DEAN; GHEMAWAT, 2010b); (HASHEM et al., 2015b).

2.3 AVALIAÇÃO DE DESEMPENHO

A avaliação de desempenho tornou-se um pré-requisito para cada estágio da vida de um sistema computacional, desde a criação do projeto até sua fabricação, além de um possível aprimoramento futuro (JAIN, 1991). Em avaliação de desempenho de sistemas, três técnicas são bastante abordadas, são elas: medição, modelagem e simulação (MENASCE et al., 2004); (LILJA, 2005); (FEITELSON, 2015). A medição consiste em obter métricas de desempenho através de experimentos controlados em um ambiente real. Ferramentas como *SYSSTAT* (JUVE et al., 2015) e *PERFMON* (WINDOWS, 2022) captam informações importantes sobre processadores, memória e outros recursos de máquinas virtuais ou físicas. O tempo de amostragem e o intervalo de medição devem ser definidos de forma que os

experimentos sejam consistentes. Na técnica de modelagem, os componentes e recursos do sistema são representados através de elementos gráficos aliados à formalismos matemáticos, como redes de filas, cadeias de *Markov* (KIM et al., 2009). Por fim, a técnica de simulação é adotada quando pesquisadores pretendem obter uma visualização da dinâmica do processo analisado que estejam ainda em sua fase de protótipo. Em computação, a simulação geralmente é abordada para eventos discretos (KIM et al., 2009).

2.4 AVALIAÇÃO DE DISPONIBILIDADE

A disponibilidade de um sistema é a probabilidade de que ele esteja operacional durante um determinado período de tempo, ou tenha sido restaurado após a ocorrência de um defeito.

Uptime é o período de tempo em que o sistema está operacional, downtime é o período de tempo em que o sistema não está operacional devido à ocorrência de um defeito ou atividade de reparo, e o $uptime + downtime$ é o período de tempo de observação do sistema. A Equação 2.1 representa a disponibilidade de um sistema (XIE et al., 2004); (KUO; ZUO, 2003); (RUPE, 2003).

$$A = \frac{uptime}{uptime + downtime} \tag{2.1}$$

2.5 AVALIAÇÃO DE PERFORMABILIDADE

A avaliação de performabilidade descreve o efeito de eventos de falhas e atividades de reparo na degradação do desempenho de sistemas. Para a avaliação de performabilidade é comum a utilização de técnicas de modelagem hierárquicas para combinação de um modelo de dependabilidade de alto nível e modelos de desempenho de baixo nível, um modelo de desempenho para cada estado do modelo de dependabilidade (PULIAFITO A.; RICCOBENE, 1996). A dependabilidade é a capacidade que um sistema tem de oferecer um serviçoo de forma confiável. A integração da modelagem de aspectos de desempenho e dependabilidade de sistemas é conhecida como modelagem de performabilidade. A modelagem de performabilidade permite a avaliação de desempenho considerando a degradação dos níveis de serviço provocados pelos eventos de falhas durante um determinado período de tempo (SAHNER R. A.; TRIVEDI, 1996).

2.5.1 REDES DE PETRI ESTOCÁSTICAS

Redes de *Petri* é uma técnica de representação matemática que pode modelar sistemas distribuídos e permite avaliar a estrutura e o comportamento do sistema modelado. Uma rede de *Petri* R é uma quíntupla R = (P, T, I, O, K), onde P = {p1,p2,...,pn} é um conjunto finito não-vazio de lugares, T = {t1, t2,..., tm} é um conjunto finito não-vazio de transições. . I : T → P é um conjunto de bags que representa o mapeamento de transições para lugares de entrada. O : T → P é um conjunto de bags que representa o mapeamento de transições para lugares de saída. K : P → N é o conjunto da capacidades associadas a cada lugar, podendo assumir um valor infinito. A representação da rede de *Petri* é formada por dois componentes: um ativo chamado de transição (seta) e outro passivo denominado de lugar (círculo). Os lugares equivalem às variáveis de estado e as transições correspondem às ações realizadas pelo sistema. Esses dois componentes são ligados entre si através de arcos dirigidos. Os arcos podem ser únicos ou múltiplos (MOLLOY, 1982). Os elementos que constituem uma rede de *Petri* são demonstrados na Figura 1, os quais serão explanados a seguir.

Figura 1 – ELEMENTOS DE UMA REDE DE PETRI - FONTE: AUTOR

Em geral, uma rede de *Petri* é um grafo bipartido dirigido, o qual consiste em dois tipos de nós, denominados de lugar, e transição. Graficamente, lugares são representados por um círculo ou elipse (REISIG, 2014), e são associados a um componente passivo destinado a retratar uma condição ou armazenar objetos (BAUSE; KRITZINGER., 2002). A mudança das condições de um sistema, o que pode ser visto também como uma mudança

de valores, é representada pelas transições, as quais são simbolizadas por um retângulo e são caracterizadas como o componente ativo das redes de *Petri* (REISIG, 2014). Por ser um grafo bipartido, a conexão entre os elementos deve ser feita considerando os dois tipos de nós, isto é, um lugar pode apenas se conectar com uma transição e vice-versa (BAUSE; KRITZINGER., 2002). Lugares e transições são conectados diretamente através de arcos (arcs). Graficamente, um arco é representado por uma seta e não constitui um componente do sistema, mas apenas uma abstrata relação, como por exemplo, conexões lógicas (REISIG, 2014). Existem dois tipos de arcos: arcos de entrada (*input arcs*) e arcos de saída (output arcs). Arcos de entrada efetuam a conexão de um lugar de entrada (*input place*) para uma transição, enquanto que arcos de saída conectam uma transição para um lugar de saída (*output place*) (BOLCH, 2006).

As redes de *Petri* estocásticas (SPN) são compostas por transições imediatas e transições temporizadas. Tempos de disparos aleatórios distribuídos exponencialmente são associados as transições temporizadas e as transições imediatas disparam com tempos iguais a zero, com prioridade superior a transições temporizadas (MOLLOY, 1982).

As redes de *Petri* estocástica são uma extensão das redes de Petri utilizada para a modelagem de desempenho e dependabilidade (TORRES E.; CALLOU, 2018). Nas SPN são adicionados dois elementos, como pode ser visto na Figura 2, que são as transições temporizadas (Figura 2 (a)) e o arco inibidor (Figura 2 (b)). A transição temporizada utiliza tempos associados, de modo que o período de habilitação da transição corresponde ao período de execução da atividade. Já o arco inibidor desativa a transição se houver tokens no lugar de origem do arco inibidor, e vai ser ativada quando não houver tokens (SILVA, 2016).

2.5.2 TÉCNICAS DE APROXIMAÇÃO DE FASES

A técnica de aproximação de fases pode ser aplicada para modelar ações, atividades e eventos não-exponenciais através do moment matching. Esta técnica indica a distribuição de probabilidade expolinomial que melhor representa a distribuição de probabilidade empírica dos tempos de análise do data set analisado. O método apresentado calcula o primeiro momento em torno da origem (média) e o segundo momento central (variância) e estima os momentos respectivos da s-transição. O inverso do coeficiente de variação dos

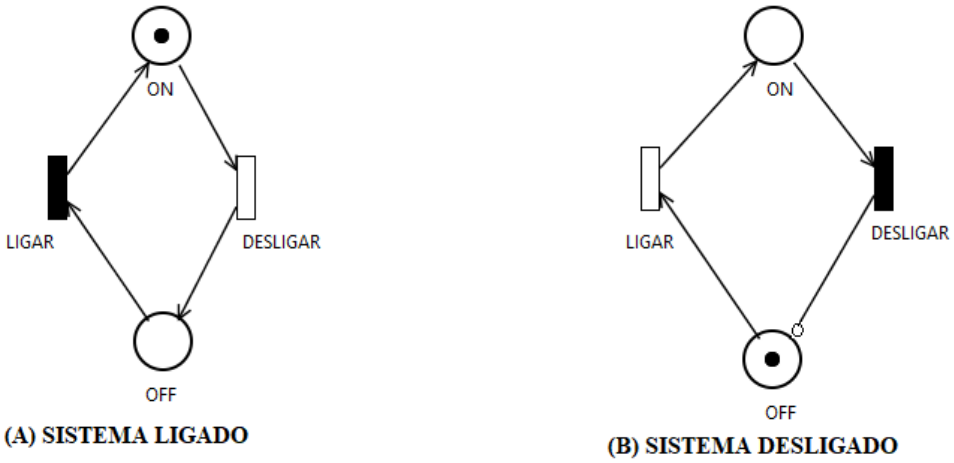

Figura 2 – EXEMPLO DE UMA REDE DE PETRI ESTOCÁSTICA - FONTE: AUTOR

dados medidos ou obtidos de um sistema permite a seleção da distribuição expolinomial que melhor se adapta à distribuição empírica. Esta distribuição empírica pode ser contínua ou discreta. Entre as distribuições contínuas, temos a Normal, Lognormal, *Weibull*, Gama, Uniforme, Pareto, Beta e Triangular e entre as distribuições discretas, temos a Geométrica, Poisson e Uniforme Discreta (YEE; VENTURA., 2000).

A técnica de aproximação de fases pode ser aplicada para modelar ações, atividades e eventos não-exponenciais através do moment matching. O método apresentado calcula o primeiro momento em torno da origem (média) e o segundo momento central (variância) e estima os momentos respectivos da s-transition. Dados de desempenho ou dependabilidade medidos ou obtidos de um sistemas (distribuição empírica) com média µ D e desvio-padrão D podem ter seu comportamento estocástico aproximados através da técnica de aproximação de fases. O inverso do coeficiente de variação dos dados medidos (Equação 2.5) permite a seleção da distribuição expolinomial que melhor se adapta a distribuição empírica (JAIN, 1991).

$$\frac{1}{CV} = \frac{\mu_D}{\sigma_D} \tag{2.2}$$

A rede de *Petri* descrita na Figura 3 representa uma atividade temporizada com distribuição de probabilidade genérica.

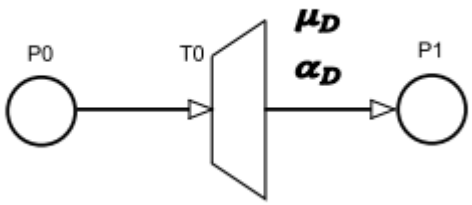

Figura 3 – Distribuição Empírica - FONTE: AUTOR

Dependendo do valor de inverso do coeficiente de variação dos dados medidos, a respectiva atividade tem uma dessas distribuições atribuídas: *Erlang*, Hipoexponencial ou Hiperexponencial.

Quando o inverso do coeficiente de variação é um número inteiro e diferente de um, os dados devem ser caracterizados através da distribuição *Erlang*, que é representada por uma sequência de transições exponenciais, cujo tamanho é calculado através da Equação 2.3. A taxa de cada transição exponencial é calculada através da Equação 2.4. Os modelos de Redes de *Petri* descritos na Figura 4 representam uma atividade temporizada com comportamento definido por uma distribuição de probabilidade *Erlang*.

$$\gamma = (\frac{\mu}{\sigma})^2 \tag{2.3}$$

$$\lambda = \frac{\gamma}{\mu} \tag{2.4}$$

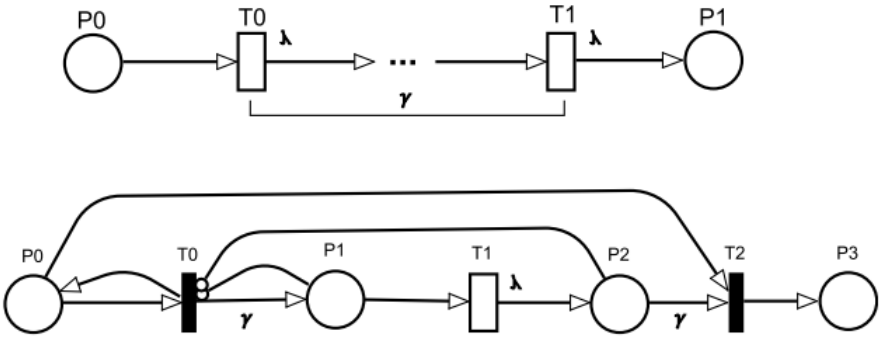

Figura 4 – Distribuição Erlang - FONTE AUTOR

Quando o inverso do coeficiente de variação é um número maior que um (mas não é um número inteiro), os dados são representados através da distribuição hipoexponencial, a qual é representada por uma sequência de transições exponenciais, cujo tamanho é calculado através da Equação 2.5. As taxas das transições exponenciais são calculadas

através das Equações 2.6 e 2.7, e os tempos médios atribuídos às transições exponenciais são calculados através das Equações 2.8 e 2.9. Os modelos de Redes de *Petri* apresentados na Figura 5 descrevem uma atividade temporizada com comportamento definido por uma distribuição de probabilidade hipoexponencial.

$$(\frac{\mu}{\sigma})^2 - 1 \leq \gamma < (\frac{\mu}{\sigma})^2 \tag{2.5}$$

$$\lambda_1 = \frac{1}{\mu_1} \tag{2.6}$$

$$\lambda_2 = \frac{1}{\mu_2} \tag{2.7}$$

$$\mu_1 = \mu \mp \frac{\sqrt{\gamma(\gamma+1)\sigma^2 - \gamma\mu^2}}{\gamma+1} \tag{2.8}$$

$$\mu_2 = \gamma\mu \pm \frac{\sqrt{\gamma(\gamma+1)\sigma^2 - \gamma\mu^2}}{\gamma+1} \tag{2.9}$$

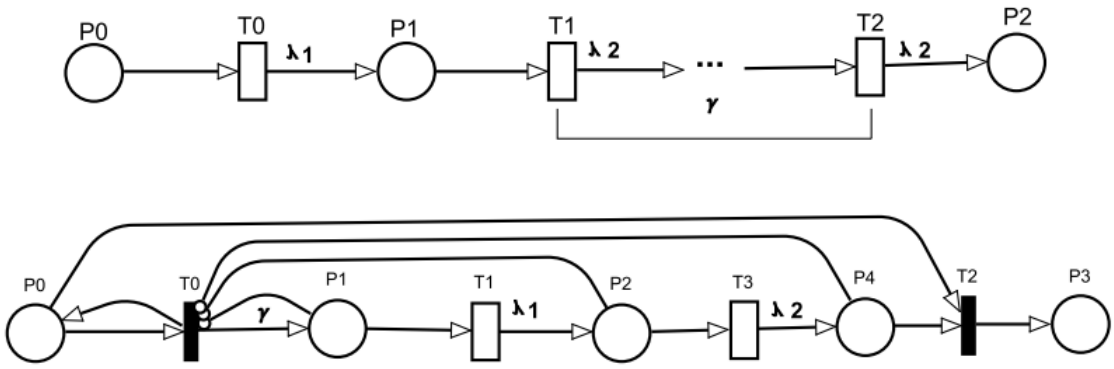

Figura 5 – Distribuição Hipoexponencial

Quando o inverso do coeficiente de variação é um número menor que um, os dados devem ser caracterizados através de uma distribuição hiperexponencial. A taxa da transição exponencial deve ser calculada através da Equação 2.10, e os pesos das transições imediatas são calculados através das Equações 2.11 e 2.12. O modelo de Redes de *Petri* que representa uma atividade temporizada com comportamento definido por uma distribuição de probabilidade hiperexponencial é descrito na Figura 6.

$$\lambda_h = \frac{2\mu}{\mu^2 + \sigma^2} \tag{2.10}$$

$$r_1 = \frac{2\mu^2}{\mu^2 + \sigma^2} \tag{2.11}$$

$$r_2 = 1 - r_1 \tag{2.12}$$

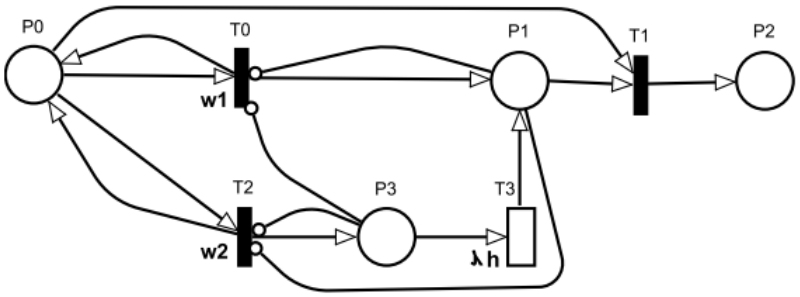

Figura 6 – Distribuição Hiperexponencial - FONTE AUTOR

2.6 DIAGRAMAS DE BLOCO DE CONFIABILIDADE

O RBD (*Reliability Block Diagram*) é uma representação da estrutura lógica de um sistema por meio de blocos ligados de acordo com suas funções ou uma relação de confiabilidade. Este é considerado um modelo do tipo combinatorial não baseado em estados. Na análise de confiabilidade do sistema, estamos interessados em encontrar a relação entre confiabilidade de componentes e confiabilidade do sistema (KUO; ZUO, 2003). Essa configuração pode ser simples como unidades dispostas em série ou paralelo. Podem também existir sistemas de configurações combinadas série/paralelo ou sistemas complexos que não podem ser decompostos em grupos de configurações em série e em paralelo (RELIAWIKI, 2017).

Um diagrama que possui componentes conectados com uma estrutura em série requer que cada componente esteja funcionando para que o sistema torne-se operacional, ou seja, a falha de qualquer componente faz com o que o sistema falhe. Quando o diagrama possui componentes ligados em paralelo, é necessário que pelo menos um dos componentes funcione, portanto, somente a falha de todos os componentes faz com que o sistema falhe. Uma forma de calcular a disponibilidade do sistema é subdividi-lo em subsistemas e calcular a disponibilidade individualmente. A disponibilidade A_S de um sistema com blocos em série é representado pela Equação 17, onde A_i é a disponibilidade de um componente i. Em um sistema com blocos paralelos simples, pelo menos uma unidade deve ter sucesso para que o sistema tenha êxito e sua disponibilidade A_p é dada pela Equação 2.18.

$$A_s = \prod_{i=1}^{n} A_i \tag{2.13}$$

$$A_p = 1 - \prod_{i=1}^{n} (1 - (A_i)) \tag{2.14}$$

A Figura 7 ilustra um RBD com três componentes conectados em série e a Equação 2.19 calcula a sua disponibilidade. A Figura 8 mostra um RBD com três componentes em paralelo e sua disponibilidade A_p é obtida por meio da equação 2.20.

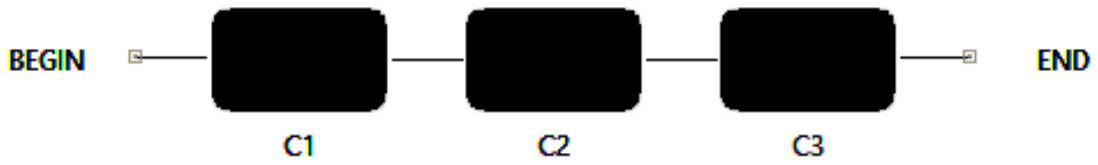

Figura 7 – RBD COM TRÊS COMPONENTES EM SÉRIE

$$A_s = A_{c1} X A_{c2} X A_{c3} \tag{2.15}$$

$$A_p = 1 - (1 - A_{c1}) X (1 - A_{c2}) X (1 - A_{c3}) \tag{2.16}$$

2.7 PLANEJAMENTO DE EXPERIMENTOS

O planejamento dos experimentos adota uma técnica de planejamento que possibilita o estudo simultâneo dos fatores considerados nos experimentos (GREMYR et al., 2003).

Existe três tipos de planejamentos de experimentos: Planejamento Simples, planejamento completo e planejamento fatorial fracionado.

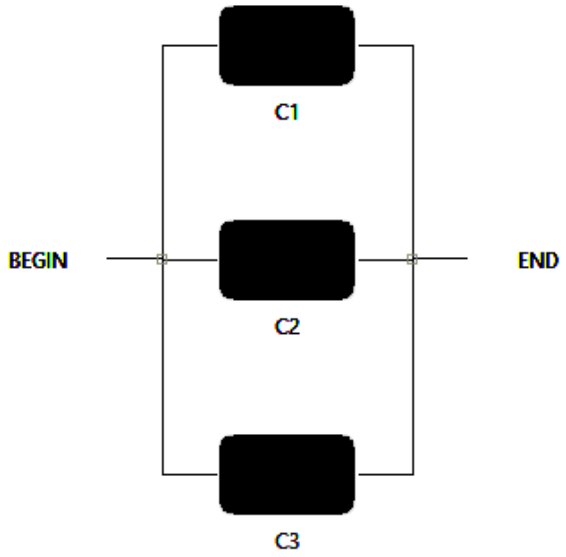

Figura 8 – RBD COM TRÊS COMPONENTES EM PARALELO

O Planejmamento Simples: É aquele que possui o modelo mais simples de um experimento fatorial, apenas uma observação para cada combinação dos dois fatores envolvidos, ou seja, não existem repetições.

O Planejamento Fatorial Completo: É uma técnica bastante utilizada quando se tem duas ou mais variáveis independentes (fatores). Ele permite uma combinação de todas as variáveis em todos os níveis, obtendo-se assim uma análise de uma variável, sujeita a todas as combinações das demais. Planejamento fatoriais são extremamente úteis para medir os efeitos (ou influências) de uma ou mais variáveis na resposta de um processo.

O planejamento Fatorial Fracionado, diferentemente de um planejamento fatorial completo, por razões de economia, limita a coleta de dados a um subconjunto das possíveis combinações dos fatores. Eles são importantes em situações reais nas quais existem muitos fatores.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta os trabalhos relacionados à avaliação de desempenho, de dependabilidade e performabilide de ambientes big data na nuvem privada. Essa seção apresenta trabalhos relacionados à avaliação de ambientes big data ou computação em nuvem. Em seguida, trabalhos sobre avaliação de dependabilidade de ambientes *big data* ou computação em nuvem são mostrados. Depois, trabalhos sobre performabilidade de ambientes *big data* ou computação em nuvem são mostrados. Finalmente, este capítulo apresenta uma comparação entre os trabalhos correlatos e o trabalho proposto.

A pesquisa pelos trabalhos relacionados ocorreu nas bases científicas IEEE e ACM, considerando o período de 5 anos. Além disso, para a pesquisa dos trabalhos relacionados foram adotadas as *strings* de busca *((Performance evaluation) and ((cloud computing) or (big data))), ((depebdability evaluation) and ((cloud computing) or (big data))) e ((performability evaluation) and ((cloud computing) or (big data)))* com o próposito de contemplar as áreas de estudo abordadas.

3.1 AVALIAÇÃO DE DESEMPENHO

O trabalho de FÈ (FÉ, 2017) propõe modelos em SPN e de otimização para auxiliar no planejamento de sistemas de transcodificação de vídeo em nuvem privada e pública. Os modelos em SPN propostos foram adotados para o cálculo das métricas de vazão, tempo de resposta e custo. Experimentos realizados testes no ambiente de nuvem *Cloudstack*, com o *JMeter* enviando requisições de transcodificações resultaram na análise do aumento de custo de aproximadamente 300% quando há uma redução no tempo de resposta de 30 para 15. Mas quando há uma redução no tempo de resposta de 45 para 30 segundos o aumento do custo financeiro será de 6%.

O trabalho de LI (LI; LIN, 2017) propôs um modelo de filas para simulação do desempenho da infraestrutura de nuvem. O experimento utilizou duas máquinas com configurações diferentes, uma do tipo *small* com 1GB de memória e outra do tipo *large* com 4GB de memória. As duas configurações de máquina virtual apresentavam uma

CPU contendo 16 núcleos. Os experimentos proporcionaram a obtenção da utilização de recursos para determinado número de *Hosts*.

O artigo de BERTONCELLO (BERTONCELLO, 2018) apresentou a avaliação de desempenho de aplicativos *Deep Learning* (DL) no ambiente big data. A avaliação mediu o impacto do processamento de forma paralela e distribuída no *TensorFlow* DL. Os resultados apresentaram melhor desempenho no processamento distribuído, com uma aceleração de até 8x e perda de precisão menor que 5%.

No trabalho apresentado por VEIGA (VEIGA et al., 2016) foi avaliado o desempenho do *Haddop cluster* através da medição da métrica tempo de execução. Como uma extensão deste trabalho, apresentaram a avaliação de desempenho do *Hadoop, Spark e Flink* por meio do *benchmark Terasort*. Os resultados mostraram que substituir o *Hadoop* pelo *Spark ou Flink* pode melhorar significativamente o desempenho no processamento de dados.

No trabalho apresentado por ECKROTH (ECKROTH, 2016) ocorreu a avaliação de custo do Hadoop suportado por um ambiente de nuvem computacional. Os fatores de análise utilizados para o estudo foram a oferta de serviço e a quantidade de data nodes para a análise Big Data. A métrica de resposta condicionada à estes fatores foram o tempo do job em minutos(tempo de execução). A oferta de serviço contribuiu de forma significativa para o tempo de execução de uma análise *big data*, mas quando o número de data nodes aumenta, o impacto se apresenta mais significativo nesta métrica. Adicionalmente, o custo foi avaliado de acordo com a plataforma de nuvem, considerando uma carga de trabalho de 37 GB e 10 data nodes no *cluster*.

Em (MACHADO, 2018a) mostrou experimentos utilizando diferentes configurações na utilização de recursos para um determinado hosts. Essa pesquisa utiliza um modelo SPN representados através de redes de *Petri* estocástica e pode ser utilizado em várias infraestrutura de nuvem. O resultado da avaliação dos modelos concebidos para representação dos cenários permite a análise de cenários que atendam aos requisitos de desempenho.

3.2 AVALIAÇÃO DE DEPENDABILIDADE

Os autores (MELO, 2017) propõem modelos RBD e SPN para a avaliação da disponibilidade orientada à capacidade de nós em uma nuvem privada *Eucalyptus*. Estes cálculos apontam quantas máquinas virtuais estarão disponíveis em um nó. Também descrevem a quantidade de recursos que serão perdidos devido à ocorrência de falhas e reparos. No mesmo estudo, (MELO, 2017) demonstram que os modelos RBD são utilizados para cálculo de métricas como disponibilidade estacionária e downtime anual.
Já o modelo SPN é utilizado para calcular a disponibilidade orientada à capacidade. Esta dissertação utiliza algumas métricas do trabalho de (MELO, 2017), assim como alguns de seus MTTF e MTTR de componentes de *hardware* e componentes da nuvem.

O trabalho (DANTAS, 2018) propôs modelos de desempenho, disponibilidade, custo e um mecanismo para avaliação de espaço de projeto para infraestrutura de nuvem com um suporte (sistema de cluster integrado ao *front-end*) a um serviço de vídeo para um subsistema de cluster em uma máquina física independente. Foi utilizado como o gerenciador de nuvem o ambiente textitEucalyptus. Em seu seu estudo de desempenho foi utilizado o *JMeter* para geração de carga e para validação do modelo proposto. E para calcular a disponibilidade foi calculado o MTTF e MTTR do *hardware* e sistema operacional, a disponibilidade do sistema aumenta significativamente de 98% no cenário com um *cluster* físico independente para 99% no cenário em que temos um *Front-end* e um *cluster* juntos em um mesmo meio físico

O trabalho (LIU B.; CHANG, 2018) aplicou técnicas de modelagem analítica e análise de sensibilidade para investigar a disponibilidade de infraestruturas como serviço em data center considerando políticas de reparo. As políticas de reparo foram modeladas através das redes de *Petri* de Recompensa, do inglês, SRN. Um conjunto de máquinas físicas representado pelo modelo SRN, e associado a políticas de reparo, analisou o impacto da disponibilidade na infraestrutura como serviço.

O trabalho (TORRES E.; CALLOU, 2018) apresentou uma abordagem para avaliar a disponibilidade e o desempenho (vazão de arquivos transferidos) de um serviço de armazenamento de dados hospedado em uma nuvem privada por meio de modelos analíticos. Os autores utilizaram uma estratégia de modelagem hierárquica através de modelos RBD, CTMC e SPN. Os modelos foram usados para avaliar a disponibilidade do serviço da

nuvem privada usando vários níveis de redundância entre os componentes. Além disso, os autores validaram o modelo de desempenho considerado um experimento com medições. Nos trabalhos mencionados, o foco da avaliação dos serviços computacionais volta-se à disponilididade, ainda que na modelagem adotada poderiam ter sido abordadas análises quanto à confiabilidade e ao custo.

O artigo (VASCONCELOS, 2019) propôs três modelos SPN representando arquiteturas de transcodificação de vídeo com diferentes números de nós, buscando o aumento da disponibilidade. O primeiro modelo representou a arquitetura baseline que dispõe de um nó, sua disponibilidade foi calculada juntamente com sua disponibilidade orientada à capacidade e disponibilidade de capacidade total. Em seguida, efetuou uma análise de sensibilidade e detectado componentes que mais influencia na disponibilidade. A representação deste primeiro modelo ocorreu através da modelagem hierárquica, utilizando como base modelos RBD para modelar subsistemas envolvidos e extrair as métricas de MTTF e MTTR utilizadas por este e subsequentes modelos. o primeiro modelo apresentou uma disponibilidade de 98.9859%, o segundo modelo mostrou a disponibilidade em 99.8807%, um ganho possível através da redundância dupla do nó e o terceiro modelo apresentou uma disponibilizadade de 99.8991%

O trabalho (OLIVEIRA, 2017) propôs um *framework* para injetar e monitorar falhas em nuvem computacional. Este *framework* busca diminuir o retrabalho por parte dos desenvolvedores que estudam métricas de disponibilidade, a fim de tornar-se uma ferramenta capaz de ser adaptável a qualquer sistema de nuvem no qual for utilizada. Após a utilização do framework o resultado encontrado após a injeção de falhas foi de 95,59% e a disponibilidade do modelo antes da injeção foi de 96,27%.

O trabalho proposto, no entanto, combina modelos RBD de baixo nível e modelos RBD e alto nível conforme aos componentes da nuvem computacional. Diferente do trabalho proposto nesta dissertação, nenhum dos trabalhos relacionados apresenta modelos para calcular o MTTF e MTTR dos componentes e a disponibilidade do ambiente *biga data* na nuvem privada.

3.3 AVALIAÇÃO DE PERFORMABILIDADE

O trabalho (EVER, 2017) realizou uma análise de performabilidade de uma infraestrutura como serviço em nuvem utilizando modelos de filas e modelos de cadeias de *Markov* de tempo contínuo (CTMC). O estudo apresentado modela o comportamento do ambiente por meio das métricas como tamanho médio da fila, vazão e tempo de resposta. Foi considerada uma abordagem de computação em nuvem em larga escala, com cerca de 5.000 servidores. A avaliação de performabilidade considerou a variação na taxa de reparo no servidor. A partir do modelo, é possível realizar um planejamento de capacidade, identificar gargalos e analisar a disponibilidade da infraestrutura. Contudo, os autores não contemplaram tais resultados.

O trabalho (ARAÚJO, 2019) propôs uma estratégia de modelagem hierárquica, que considera as vantagens das redes de *Petri* estocásticas (SPN) e diagramas de blocos de confiabilidade (RBD) para avaliar a disponibilidade, disponibilidade orientada a capacidade, confiabilidade e utilização de recursos das infraestruturas como serviço. Foram utilizadas métricas de performabilidade e calculadas a partir do modelo de desempenho e posteriormente combinadas com o modelo dependabilidade para ilustrar o efeito da disponibilidade no desempenho da infraestrutura. Foi possível diversificar os resultados de performabilidade, pois o usuário pôde combinar as métricas de desempenho com as métricas de dependabilidade. Permitiu a avaliação do impacto da ocorrência de eventos de falhas e atividades de reparo no desempenho da infraestrutura em nuvem por meio de métricas de performabilidade como utilização de processador e dispositivo de entrada e saída. Portanto, através dos estudos descritos, foi possível demonstrar que a metodologia adotada considerando os modelos de dependabilidade, performabilidade, custo e decisão pode auxiliar gestores e analistas no processo de tomada de decisão para escolher um conjunto de soluções.

O diferencial deste trabalho é a combinação dos resultados das avaliações de modelos de desempenho e de modelos de dependabilidade, proporcionando uma menor complexidade da avaliação da performabilidade de nuvens computacionais.

3.4 COMPARAÇÃO DOS TRABALHOS RELACIONADOS

Essa seção apresenta a comparação dos trabalhos relacionados ao estudo proposto. No estado da arte vários trabalhos abordam avaliação de desempenho, computação em nuvem, existem também trabalhos que abordam o armazenamento de banco de dados ou da infraestrutura. Destacando o diferencial dessa pesquisa e tendo suas contribuições uma metodologia para avaliação de desempenho (AVD), avaliação de dependabilidade (AD), avaliação de performabilidade (AP), Avaliação de ambientes big data (BD), Avaliação de ambientes de nuvem (N).

TRABALHO	N	AD	AVD	AP	BD
(FÉ, 2017)	X		X		
(LI;LIN, 2018)		X	X	X	X
(KASSIANO, 2019)			X		
(LI et al., 2017)			X	X	X
(BERTONCELLO, 2018)		X	X		
(VEIGA et al., 2018)		X	X		
(MELO et al,2017) X		X			
(DANTAS,2018)	X	X	X		
(LIU et al.,2018)		X	X		
(TORRES, 2018)	X	X	X		
(VASCONCELOS, 2019)	X	X	X		
(OLIVEIRA, 2017)		X		X	
(ECKROTH, 2016)		X		X	X
(EVER, 2017)		X		X	X
(ARAÚJO, 2019)	X		X	X	
ESTA DISSERTAÇÃO	X	X	X	X	X

Tabela 1 – TABELA COMPARATIVA DE TRABALHOS RELACIONADOS

Cinco trabalhos relacionados avaliaram ambientes de nuvem privada. Dez trabalhos adotaram a técnica de avaliação de dependabilidade. Onze trabalhos apresentaram avaliação de desempenho. Seis trabalhos apresentaram avaliação de performabilidade. E quatro trabalhos avaliaram ambientes *big data* na nuvem privada. E esta dissertação abordou nuvem privada, avaliação de dependabilidade, avaliação de desempenho e performabilidade.

4 METODOLOGIA PARA AVALIAÇÃO DE DESEMPENHO, DISPONIBILIDADE E PERFORMABILIDADE EM AMBIENTES BIG DATA EM NUVENS PRIVADAS

Este capítulo apresenta a metodologia proposta que é composta de quatro atividadess. Inicialmente a atividade proposta para avaliação de desempenho de ambientes *big data* na nuvem será apresentado. Em seguida, a atividade para avaliação de disponibilidade de ambientes big data na nuvem privada será mostrado. Finalmente, será mostrado a atividade para avaliação do impacto da disponibilidade no desempenho de ambientes *big data* nuvem privada.

4.1 VISÃO GERAL DA METODOLOGIA PROPOSTA

A metodologia proposta para avaliação de desempenho, disponibilidade e performabilidade de ambiente *big data* na nuvem privada pode ser adotada por profissionais de TI como engenheiro de software, arquiteto de sistemas e infraestrutura com conhecimento em computação em nuvem, técnicas de planejamento de experimentos, redes de Petri estocásticas e diagramas de blocos de confiabilidade podem utilizar a metodologia proposta. Esta metodologia é composta das atividades para avaliação de desempenho de ambientes *big data* na nuvem privada, método para avaliação de disponibilidade de ambientes big data na nuvem privada e atividade para avaliação de performabilidade de ambientes *big data* na nuvem privada.

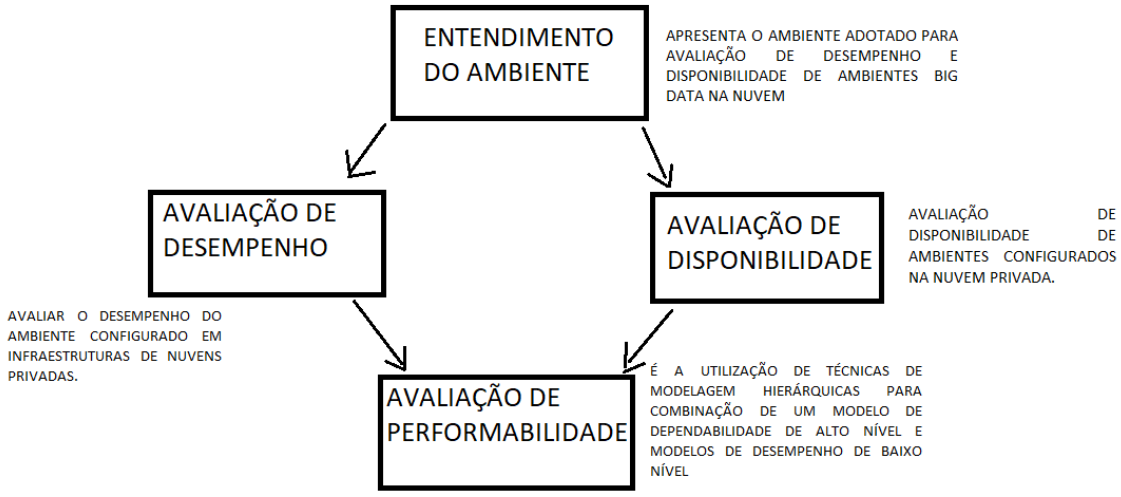

Figura 9 – VISAO GERAL

4.2 ATIVIDADE PARA AVALIAÇÃO DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

A atividade proposta é composta, conforme Figura 1: entendimento e configuração do ambiente *big data* na nuvem privada, planejamento de experimentos do ambiente big data na nuvem privada, geração de carga de trabalho de dados de redes sociais, modelagem de desempenho de ambientes *big data* na nuvem privada, medição de desempenho de ambientes *big data* na nuvem privada, análise estatística de métricas de desempenho de ambientes big data na nuvem privada, refinamento do modelo de desempenho de ambientes *big data* na nuvem privada, mapeamento das métricas de desempenho, validação do modelo de desempenho de ambientes *big data* na nuvem privada, e análise de novos cenários de ambientes *big data* na nuvem privada.

4.2.1 ENTENDIMENTO, OBJETIVOS E CONFIGURAÇÃO DO AMBIENTE BIG DATA NA NUVEM PRIVADA

Para avaliar o desempenho do ambiente *big data* configurado em infraestruturas de nuvens privadas teve como objetivos necessários entender os requisitos dos serviços configurados. Em seguida, com base nesses requisitos foram escolhidas a plataforma de nuvem e a aplicação *big data*. Essa atividade também considerou a identificação das métricas para avaliação de desempenho da aplicação *big data* na nuvem privada.

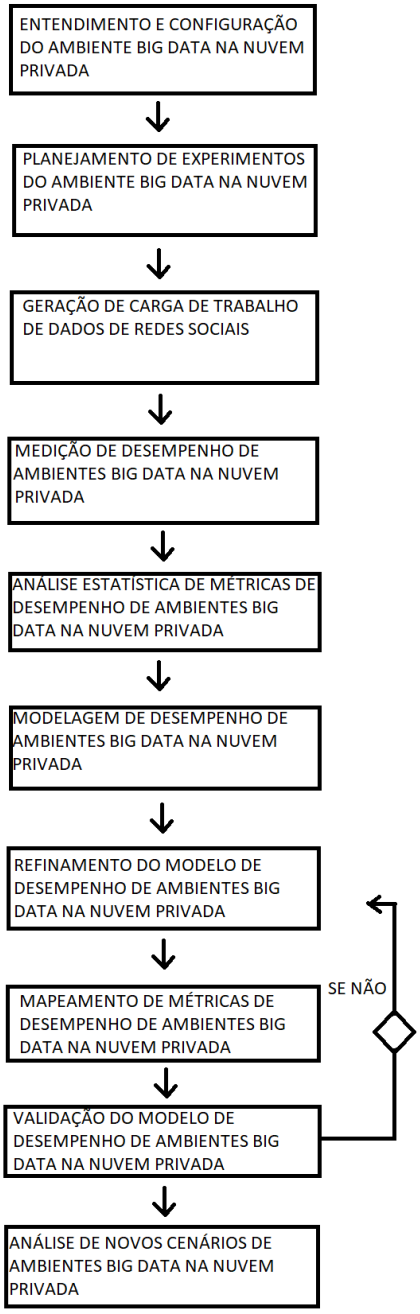

Figura 10 – ATIVIDADE PARA AVALIAÇÃO DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

A plataforma de nuvem privada escolhida deve ser configurada considerando máquinas virtuais com diferentes ofertas de serviço. As principais plataformas de nuvem privadas que podem ser adotadas são o *Apache Cloudstack* (CLOUDSTACK, 2022), o *Apache OpenStack* (OPENSTACK, 2022) e o *Eucalyptus* (OPENSTACK, 2022). Da mesma forma, o ambiente *big data hadoop cluster* (HADOOP, 2022) pôde ser configurado considerando as diferentes quantidades de *data nodes* e *master nodes*.

4.2.2 PLANEJAMENTO DE EXPERIMENTOS DO AMBIENTE BIG DATA NA NUVEM PRIVADA

No planejamento de experimentos será identificada a capacidade computacional oferecida pela nuvem privada para instanciação de máquinas virtuais. Isso permite o estabelecimento dos fatores e seus níveis. O *Hadoop cluster* é composto do *Master nodes* e *Data nodes* que são configurados na nuvem privada. Esses componentes são configurados em diferentes ofertas de serviço da infraestrutura de nuvem. A quantidade de *Master nodes* e data nodes depende da oferta de serviço provida pela nuvem privada. Desta forma, a oferta de serviço e o número de *data nodes* podem ser considerados fatores do planejamentos de experimentos e suas variações os níveis desses fatores. Desta forma, *small, medium e large* seriam os níveis da oferta de serviço e 2, 4 e 6 seriam os níveis do número de *data nodes*.

4.2.3 GERAÇÃO DE CARGA DE TRABALHO DE DADOS DE REDES SOCIAIS

A atividade de geração de carga de trabalho proporciona dados de redes sociais que serão analisados no ambiente *big data* configurado na nuvem privada. O conjunto de dados pode ser capturado das redes sociais, através de ferramentas de *software* como o *RStudio* (RSTUDIO, 2022) e um algoritmo de captura de dados para fins de análise estatística. Esse conjunto de dados pode ser processado e analisado por softwares como o *MapReduce* e o *Apache Spark* (HADOOP, 2022).

4.2.4 MEDIÇÃO DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Essa atividade proporciona a medição das métricas de desempenho selecionadas, considerando uma determinada configuração do ambiente *big data* na nuvem privada e uma carga de trabalho. A medição das métricas selecionadas ocorre em cada experimento planejado, por pelo menos 30 vezes. Métricas como o tempo de execução (seg) e a utilização de recursos (%) podem serem adotadas para avaliação de desempenho de ambientes *big data* na nuvem privada. Nessa atividade também são definidos os intervalos de medição e de amostragem para as coletas das métricas. Ferramentas como *SYSSTAT* (CENTOS, 2022) e *PERFMON* (WINDOWS, 2022) proporcionam a coleta das métricas de desempenho como utilização de recursos e tempo de resposta. Além disso, o ambiente de nuvem privada e as máquinas virtuais devem ser reinicializadas após cada experimento para evitar alteração dos resultados das métricas. Ao armazenar o log com os dados das métricas de desempenho, verifica-se a existência de alguma inconsistência com os dados medidos. Caso haja, uma nova medição de desempenho será realizada. Caso contrário, será realizada a análise estatística das métricas de desempenho.

4.2.5 ANÁLISE ESTATÍSTICA DE MÉTRICAS DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Esta atividade tem como objetivo a análise estatística das métricas de desempenho escolhidas. O resultado dessa análise é o cálculo das médias e desvios-padrões das métricas de desempenho adotadas referentes à cada cenário configurado de acordo com os fatores e níveis definidos no planejamento de experimento mostrado na Tabela 13. Além disso, é realizada a análise da existência de outliers que podem ter sido causados por erros menores tais como perturbações no ambiente de medição através da ferramenta Minitab (MINITAB, 2022).

4.2.6 MODELAGEM DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Aplicações realizadas em ambientes de nuvem podem ser modeladas por redes de Petri estocásticas, em que modelos cliente/servidor (JAIN, 1991) se comunicam através de interfaces de rede da nuvem. Métricas importantes para a qualidade e gerenciamento de aplicações realizadas em ambientes de nuvem podem ser extraídas de modelos baseados em redes de Petri estocásticas (MARINESCU, 2017) .

A modelagem de desempenho proposta baseada em SPN (CENTOS, 2022) e considera a avaliação do desempenho do *Hadoop cluster* configurado na nuvem privada. O modelo SPN proposto na (Seção 4.1) representa o *hadoop cluster* e a carga de trabalho através de duas sub-redes. A modelagem de desempenho proposta considera Hadoop clusters compostos por master node e data nodes. O master node coordena e gerencia recursos do cluster, e os data nodes analisam data sets que podem ser gerados por diferentes fontes, tais como redes sociais, dados meteorológicos e de saúde.

4.2.7 REFINAMENTO DO MODELO DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

O refinamento do modelo é realizado a partir das métricas coletadas na atividade de medição. A técnica de aproximação por fases provê a seleção da distribuição de probabilidade hipoexponencial e os parâmetros numéricos desta distribuição de probabilidade que melhor representam as métricas que foram coletadas para avaliação de desempenho do ambiente *big data* na nuvem privada (YEE; VENTURA., 2000).

4.2.8 MAPEAMENTO DE MÉTRICAS DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

O objetivo dessa atividade é representar o conjunto de critérios de desempenho de aplicações *big data* em nuvens privadas através de elementos das redes de Petri estocásticas, visto que esse formalismo matemático foi adotado para concepção do modelo proposto refinado.

4.2.9 VALIDAÇÃO DO MODELO DE DESEMPENHO DE AMBIENTES BIGDATA NA NUVEM PRIVADA

A validação do modelo de desempenho permite a comparação dos resultados das métricas de desempenho obtidas através do modelo de desempenho refinado e das métricas coletadas. A comparação dos resultados dessas métricas devem ser equivalentes a um erro de precisão aceitável. Se o valor desse erro for maior que 10%, haverá necessidade de refinar o modelo de desempenho novamente (XIE et al., 2004). Caso o erro de precisão seja igual ou menor que 10%, será realizada a análise de novos cenários. O teste pareado também pode ser adotado para avaliar quantitativamente o modelo de desempenho refinado (XIE et al., 2004).

4.2.10 ANÁLISE DE NOVOS CENÁRIOS DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Essa atividade tem como objetivo analisar novos cenários com diferentes volumes de carga de trabalho, várias configurações de máquinas virtuais e quantidades de *data nodes*. Nessa atividade, o modelo de desempenho validado é adotado para realizar a análise das métricas selecionadas, como as métricas de utilização de processador e de memória.

4.3 ATIVIDADE DE AVALIAÇÃO DE DISPONIBILIDADE DE AMBIENTES BIG DATA EM NUVEM PRIVADA

Esta seção apresenta a atividade proposta para avaliar a disponibilidade de ambientes *big data* configurados na nuvem privada. Que foram utilizados modelos baseados em RBD para representação da métrica de disponibilidade do sistema. A modelagem através de RBD foi escolhida devido a interdependência de seus componentes, para o funcionamento do sistema de forma continua por não existir prioridades entre os componentes.
A atividade proposta é composta de 4 atividades, são elas: entendimento e configuração do ambiente *big data* na nuvem privada, parametrização dos modelos de disponibilidade, geração de modelos de disponibilidade, análise de cenários de ambientes *big data* na nuvem privada.

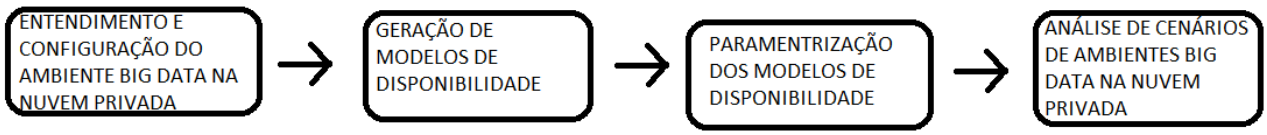

Figura 11 – MÉTODO DE AVALIAÇÃO DE DISPONIBILIDADE DE AMBIENTES BIG DATA NA NUVEM PRIVADA

4.3.1 ENTENDIMENTO E CONFIGURAÇÃO DO AMBIENTE BIG DATA NA NUVEM PRIVADA

Essa atividade tem o objetivo de avaliar a disponibilidade de ambientes *big data* em infraestruturas de nuvens considerando os diferentes serviços ofertados. Essa atividade também considera a escolha da plataforma de nuvem e do ambiente *big data*. As principais plataformas de nuvem privadas que podem ser adotadas são o *Apache Cloudstack* (CLOUDSTACK, 2022), o *Apache OpenStack* (OPENSTACK, 2022) e o Eucalyptus (EUCALYPTUS, 2022). Da mesma, forma o *Hadoop cluster* (HADOOP, 2022) pode ser configurado considerando diferentes quantidades de *data nodes e master nodes*.

4.3.2 GERAÇÃO DE MODELOS DE DISPONIBILIDADE

Esta atividade tem o objetivo de prover modelos para avaliar a disponibilidade de ambientes *big data* na nuvem privada. Dessa forma, uma estratégia de modelagem hierárquica foi proposta para combinar modelos baseados em RBD. Nessa estatégia de modelagem, modelos de baixo nível são combinados para a realização do cálculo dos parâmetros do modelo de alto nível. Os modelos de baixo nível são Modelo de *HARDWARE*, Modelo do CONTROLADOR, Modelo do CONTROLADOR *NODE*, Modelo da *VM - MASTER NODE* e Modelo da *VM - DATANODE*. O Modelo *HARDWARE* é composto do processador, memória e disco. O Modelo CONTROLADOR é composto do Modelo *Hardware*, Sistema Operacional e módulo de gerenciamento. O Modelo CONTROLADOR *NODE* é composto do *Hyper-V*, sistema operacional e o Modelo *Hardware*. O Modelo *VM-MASTER NODE* é composto do *master node*, *Hyper-V* e Sistema Operacional. O Modelo *VM-DATANODE* é composto do *DataNode*, Sistema Operacional e *Hyper-V*. E o modelo de alto nível é o MODELO DA PLATAFORMA DE NUVEM.

4.3.3 PARAMETRIZAÇÃO DOS MODELOS DE DISPONIBILIDADE

Nesta atividade serão parametrizados os modelos RBD adotados para avaliação de disponibilidade de ambientes big data na nuvem privada. Os parâmetros usados são MTTF (*Mean Time to Failure)* e MTTR *(Mean Time to Repair)* dos componentes da nuvem privada e do ambiente *big data*. Os valores desses parâmentos podem variar de acordo com o cenário adotado.

4.3.4 ANÁLISE DE CENÁRIOS DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Esta atividade tem como objetivo analisar a disponibilidade de novos cenários considerando a configuração de ambientes *big data* na nuvem privada, como por exemplo, equipamentos com diferentes valores de MTTF.

4.4 ATIVIDADE PARA AVALIAÇÃO DE PERFORMABILIDADE DE AMBIENTES BIG DATA EM NUVEM PRIVADA

Esta atividade é composta de três atividades, são elas: metódo para avaliação de desempenho de ambientes *big data* na nuvem privada, método para avaliação de disponibilidade de ambientes *big data* na nuvem privada e estratégia de composição e decomposição.

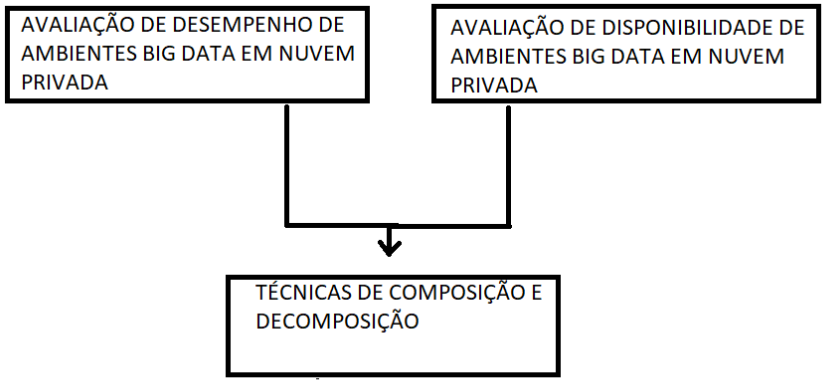

Figura 12 – MÉTODO PARA AVALIAÇÃO DE PERFORMABILIDADE DE AMBIENTES BIG DATA EM NUVEM PRIVADA

4.4.1 AVALIAÇÃO DE DESEMPENHO DE AMBIENTES BIG DATA EM NUVEM PRIVADA

Essa atividade tem o objetivo de avaliar o desempenho de ambientes *big data* com várias configurações de *master nodes e data nodes* em ofertas de serviços da nuvem privada, essas avaliações serão realizadas através de atividades de medição e modelagem. O modelo SPN proposto permite o cálculo das métricas utilização do processador e utilização de memória.

4.4.2 AVALIAÇÃO DE DISPONIBILIDADE DE AMBIENTES BIG DATA EM NUVEM PRIVADA

Essa atividade tem o objetivo de avaliar a disponibilidade de ambientes big data na nuvem. Essas avaliações serão realizadas através de modelos RBD combinados conforme a estratégia de modelagem hierárquica proposta. A análise provê a métrica de disponibilidade dos equipamentos utilizados na nuvem privada.

4.4.3 TÉCNICAS DE COMPOSIÇÃO E DECOMPOSIÇÃO

Esta atividade combina os resultados de utilização do processador e memória do modelo de disponibilidade de alto nível, o qual representa os cenários de disponibilidade, aos resultados do modelo de desempenho de baixo nível que representa os cenários de desempenho. (PULIAFITO A.; RICCOBENE, 1996).

5 MODELOS DE DESEMPENHO E DE DISPONIBILIDADE

Inicialmente, esse capítulo apresenta um modelo baseado em SPN para avaliação de desempenho de ambientes *big data* na nuvem privada. Em seguida, uma estratégia de modelagem hierárquica é mostrada para avaliação de disponibilidade desses ambientes. Finalmente, modelos baseados em RBD adotados na estatégia de modelagem proposta são apresentados.

5.1 MODELO DE DESEMPENHO

Essa seção apresenta o modelo SPN proposto para avaliação de desempenho de ambientes *big data* em infraestruturas de nuvens privadas. Os sub-modelos *Workload* e *Hadoop Cluster* do modelo de desempenho proposto representam as requisições do cliente e o ambiente *big data*, respectivamente (Figura 13). O modelo de desempenho possibilita a verificação dos vários níveis de requisições suportados pelo ambiente *big data* configurado na nuvem privada (Figura 14).

A sub-rede *Workload* representa o envio das requisições do usuário ao *Hadoop cluster* configurado na nuvem privada. A marcação (NC) atribuída ao lugar Cliente que

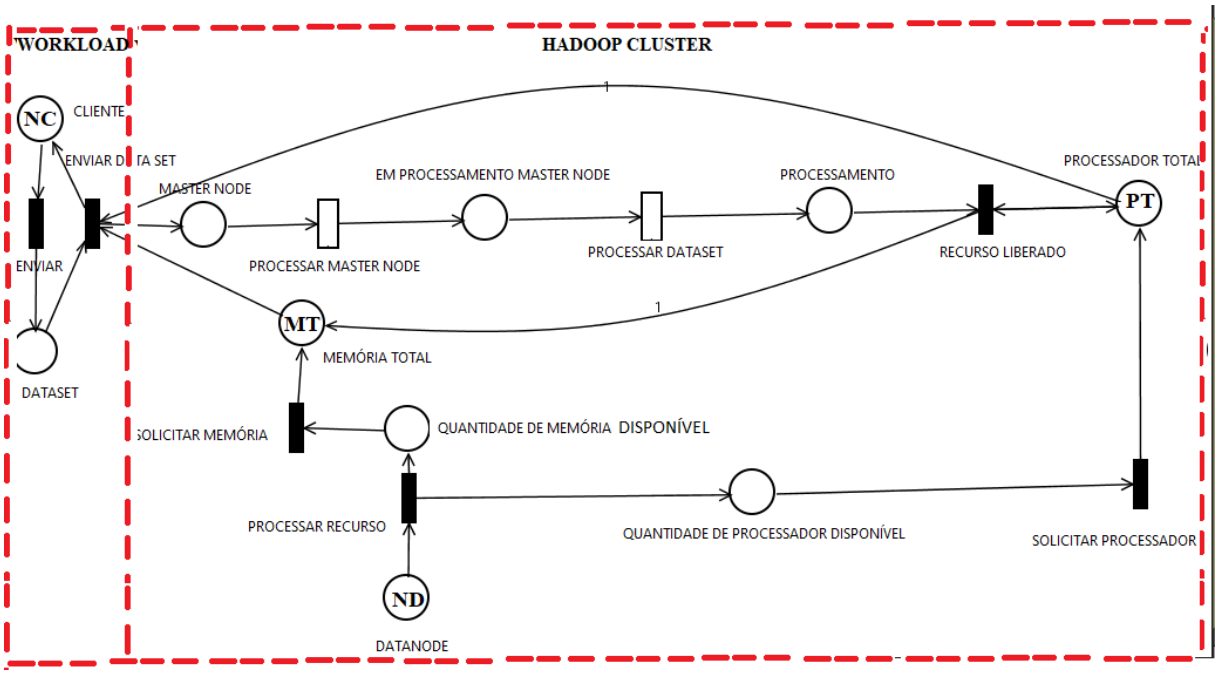

Figura 13 – MODELO DE DESEMPENHO

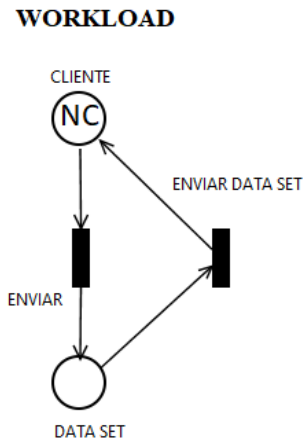

Figura 14 – SUBREDE WORKLOAD

define a carga de trabalho que será enviada ao *Hadoop cluster*, onde o número de marcações é proporcional ao tamanho do *data set*. O tempo de envio (TP) da carga de trabalho está associado a transição temporizada Enviar. Após o disparo da transição imediata Enviar, a requisição é enviada para ser atendida pelo *Hadoop cluster*. Após o disparo da transição imediata Enviar *Data SET*, a requisição é enviada ao *master node*.

A sub-rede *Hadoop Cluster* representa a infraestrutura de processamento e armazenamento da nuvem privada usada para configuração do *master node* e dos *data nodes* (Figura 14). O master node tem a função de coordenar as atividades de processamento que é representada pela transição imediata PROCESSAR *MASTER NODE*. Já o tempo (TT) associado a transição temporizada PROCESSAR DATA SET representa o tempo necessário para os data nodes do *Hadoop cluster* processarem o *dataset*. Após este tempo, o recurso de processamento e memória são liberados. A marcação (ND) associado ao lugar do *DATANODE* representa a quantidade de data nodes que compõem o *Hadoop Cluster*. As marcações (MT) e (PT) dos lugares MEMÓRIA TOTAL e PROCESSADOR TOTAL representam as capacidades da memória e processador do *Hadoop cluster*, em que a capacidade de cada data node é somada para representar a capacidade total do *cluster*. O *dataset* é processado utilizando os recursos dos data nodes. Uma vez processado o *dataset*, os recursos de processador e memória da máquina virtual são liberados.

Cada marcação atribuída ao lugar Quantidade de Processador Disponível representa a capacidade disponível de processamento na insfraestrutura instanciada na máquina virtual da nuvem privada. Cada marcação atribuída ao lugar Processador Total representa a infraestrutura de processamento total da nuvem que está sendo utilizada.

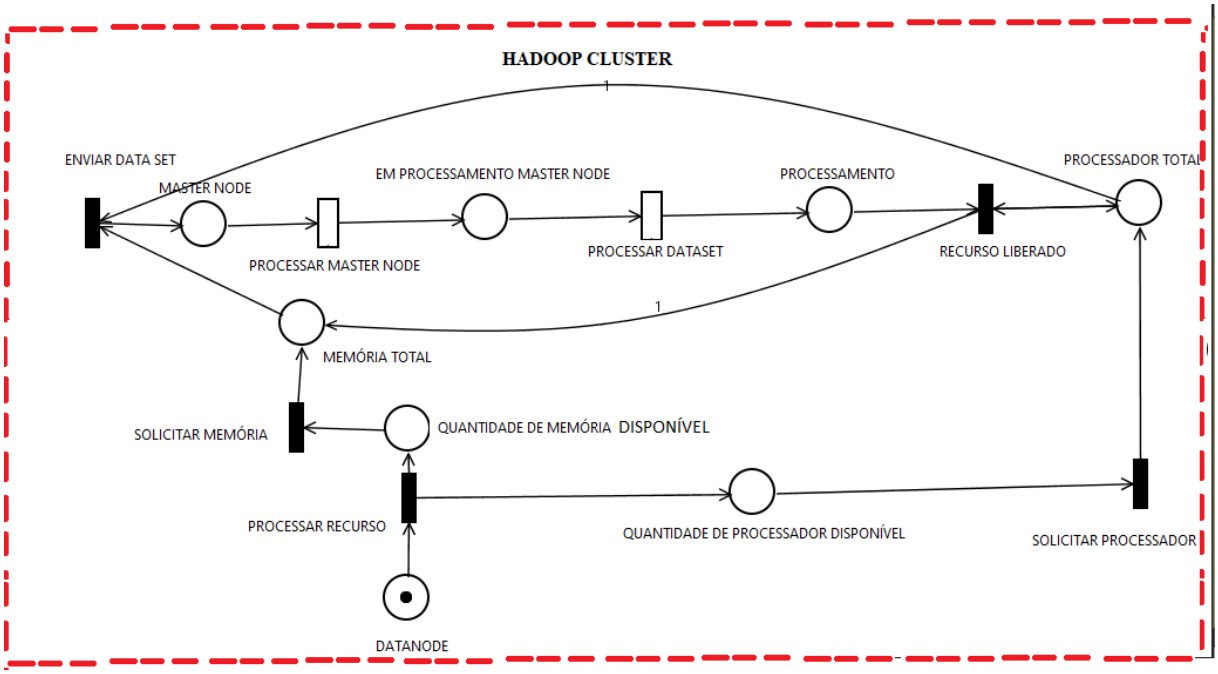

Figura 15 – SUBREDE HADOOP CLUSTER

Cada marcação atribuída ao lugar Quantidade de Memória Disponível representa a capacidade disponível de memória na insfraestrutura instanciada na máquina virtual da nuvem privada. Cada marcação atribuída ao lugar Memória Total representa a infraestrutura de memória total da nuvem que está sendo utilizada.

Esse modelo de desempenho avalia o impacto de diferentes tipos e níveis de requisições de usuários no ambiente *big data* configurado na nuvem privada através do cálculo das métricas de utilização do processador e utilização de memória. Essa avaliação proporciona meios para o planejamento de infraestruturas que atendam a determinados níveis de carga de trabalho com a qualidade de serviço desejada.

5.1.1 MÉTRICAS DE DESEMPENHO

O modelo de desempenho proporciona o cálculo das métricas de utilização de processador e de memória das infraestruturas de nuvens privadas. A utilização de memória (UM) representa a razão entre a memória usada para o atendimento das requisições dos usuários e a memória total atribuída ao *data node*. A utilização de processador (UP) representa a fração de tempo que o processador permanece ocupado atendendo às requisições feitas pelos usuários. Essa métrica é representada como porcentagem da utilização da infraestrutura de processamento total do *data node* (Tabela 2).

MÉTRICA	EXPRESSÃO
PROCESSADOR	((E{#MASTER NODE}) + (E{#EMPROCESSAMENTO MASTERNODE})+(E#PROCESSAMENTO}))X100/ (PROCESSADOR TOTAL))
MEMÓRIA	((MEMORIA TOTAL) - (E#MEMORIA TOTAL}))X100/ (MEMORIATOTAL)

Tabela 2 – MÉTRICAS DE DESEMPENHO

5.1.2 MODELO REFINADO DE DESEMPENHO

O refinamento do modelo de desempenho é realizado a partir das métricas de desempenho coletadas na atividade de medição. Essas métricas são adotadas para gerar os parâmetros do modelo de desempenho. A técnica de aproximação por fase (YEE; VENTURA., 2000) fornece a seleção da distribuição de probabilidade expolinomial e os parâmetros numéricos desta distribuição de probabilidade que melhor representam as métricas que foram coletadas para avaliação de desempenho do ambiente de *big data* em nuvem privada.

5.2 MODELOS DE DISPONIBILIDADE

Nesta seção foram utilizados modelos baseados em RBD para representação da métrica de disponibilidade do sistema. A modelagem através de RBD foi escolhida devido à interdependência de seus componentes, para o funcionamento do sistema de forma continua por não existir prioridades entre os componentes. Todos os MTTF e MTTR utilizados pelos modelos RBD são exponencialmente distribuídos. A Figura 17 mostra um modelo RBD básico que permite a representação de sistemas de alto nível e de subsistemas de baixo nível da infraestrutura da nuvem privada. Os parâmetros desse modelo são apresentados na Tabela 3. A modelagem hierárquica proporciona uma redução na complexidade da representação da nuvem privada.

A estratégia de modelagem proposta combina modelos RBD de baixo nível com um módelo RBD de alto nível. Os modelos RBD de baixo nível são adotados para representar os componentes a infraestrutura de nuvem e calcular os parâmetros do modelo de alto nível. Já o modelo de alto nível representa a infraestrutura de nuvem e provê o cálculo da disponibilidade desse ambiente (Figura 16). Os componentes MEM, PROC e DISC representam o componente do modelo RBD do *HARDWARE*, que representa um dos componentes do modelo da infraestrutura de alto nível o CONTROLADOR.

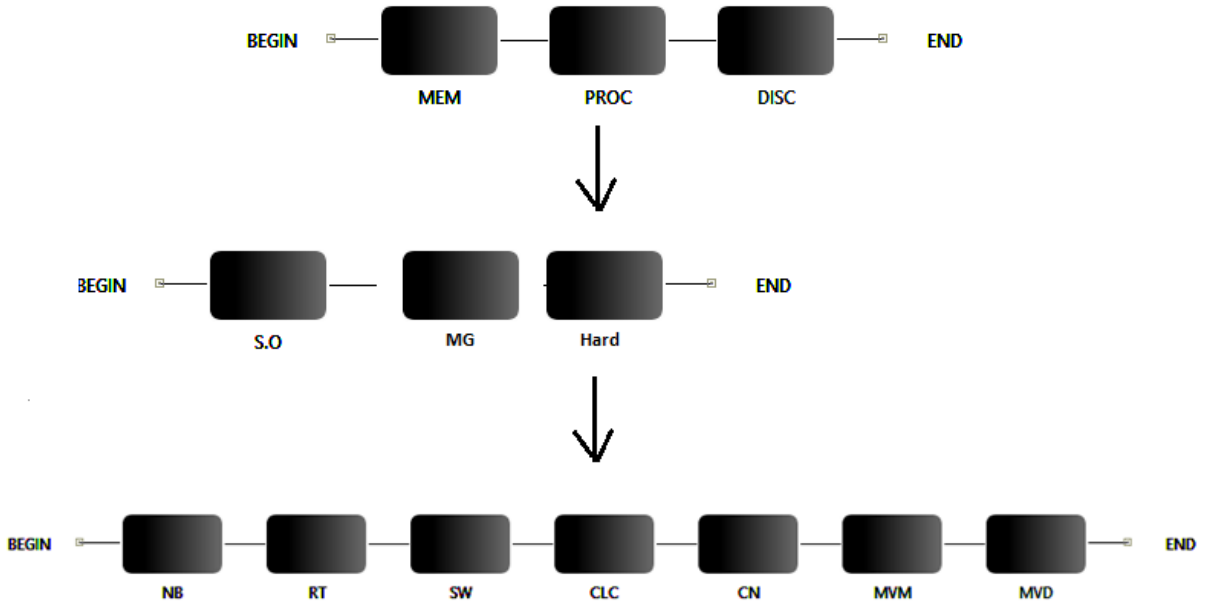

Figura 16 – DISPONIBILIDADE DO MODELO RBD

5.2.1 MODELO DA PLATAFORMA DE NUVEM

O modelo da plataforma de nuvem representa os componentes dessas plataformas através de diagramas de bloco de confiabilidade e calcula a disponibilidade da infraestrutura de nuvem por meio da disponibilidade dos seus componentes (Figura 17). A plataforma *Cloudstack* é composta pelo Controlador de Nuvem (CLC), Controlador do Node (CN), máquinas físicas onde os componentes da plataforma *CloudStack* estão configurados e conectados por meio do *nobreak, switch* e um roteador. A Tabela 3 representa os parâmetros do modelo RBD.

Figura 17 – MODELO RBD

Todos os componentes da plataforma de nuvem devem estar operacionais para que a nuvem computacional esteja operacional. O modo operacional desse ambiente de nuvem é $OM_{PN} = (NB \wedge RT \wedge SW \wedge CLC \wedge CN \wedge MVM \wedge MVD)$, onde NB, RT, SW, CLC, CN, MVM e MVD são *nobreak*, roteador, *switch*, controlador de nuvem, controlador do

PARÂMETROS	DESCRIÇÃO
$MTTF_{Block}$	Tempo Médio para Defeito do Block
$MTTR_{Block}$	Tempo Médio para Reparo do Block

Tabela 3 – PARÂMETROS DE MODELO RBD

node, máquina virtual do *master node*, máquina vitural do *data node*, respectivamente. A Figura 18 mostra o modelo RBD adotado para estimar a disponibilidade da infraestrutura dessa plataforma. Os parâmetros desse modelo são apresentados na Tabela 3.

Figura 18 – MODELO DA PLATAFORMA DE NUVEM

PARÂMETROS	DESCRIÇÃO
MTTFCLC, MTTFCN, MTTFMVM, MTTFMVD, MTTFSW, MTTFRT, MTTFNB	MTTFCLC, CN, MVM, MVD, SW, RT e NB
MTTRCLC, MTTRCN, MTTRMVM, MTTRMVD, MTTRSW, MTTRRT, MTTRNB	MTTRCLC, CN, MVM, MVD, SW, RT e NB

Tabela 4 – PARÂMETROS DO MODELO DA PLATAFORMA DE NUVEM

5.2.1.1 MODELO DO HARDWARE

O modelo RBD do *Hardware* apresentado na (Figura 19) representa os recursos de processamento e de armazenamento de um sistema computacional. O modo operacional desse modelo é $OM_{HARD} = (MEM \wedge PROC \wedge DISC)$, onde a MEM, PROC e DISC são a memória, processador e disco. Os parâmetros desse modelo são apresentados na Tabela 5.

Figura 19 – MODELO DE HARDWARE

PARÂMETROS	DESCRIÇÃO
MTTFMem, MTTFProc, MTTFDisc	Tempo Médio de Defeito de Mem, Proc, Disc
MTTRMem, MTTRProc, MTTRDisc	Tempo Médio de Reparo de Mem, Proc, Disc

Tabela 5 – PARÂMETROS DE MODELO DO HARDWARE

5.2.1.2 MODELO CONTROLADOR

O modelo RBD do CONTROLADOR (Figura 20) representa o CONTROLADOR da nuvem privada. O modo operacional desse modelo é $OM_{PN} = (S.O \wedge MG \wedge HARD)$, onde o S.O, MG e HARD são os componentes sistema operacional, módulo de gerenciamento e hardware. Salientando que uma falha em qualquer dos componentes em série acarreta na falha do sistema e na falha de provimento do serviço. Os parâmetros desse modelo são apresentados na Tabela 6.

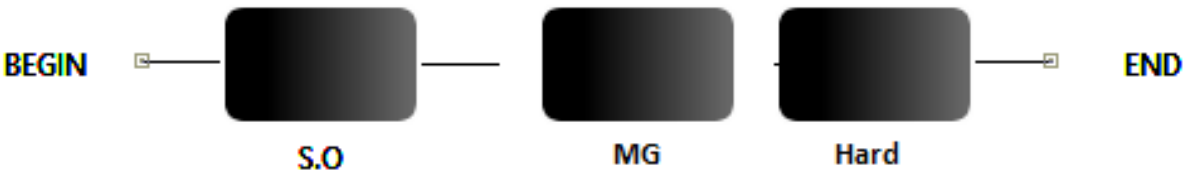

Figura 20 – MODELO DO CONTROLADOR

PARÂMETROS	DESCRIÇÃO
MTTFS.O, MTTFMG, MTTFHard	Tempo Médio de Defeito de S.O, MG, Hard
MTTRS.O, MTTRMG, MTTRHard	Tempo Médio de Reparo de S.O, MG, Hard

Tabela 6 – PARÂMETROS DE MODELO DO CONTROLADOR

5.2.1.3 MODELO CONTROLADOR NODE

O modelo RBD do CONTROLADOR NODE (Figura 21) representa o controlador do nó. O modo operacional desse modelo é $OM_{PN} = (HYPER - V \wedge S.O \wedge HARD)$, onde o HYPER-V, S.O, HARD representa componentes *Hyper-V*, sistema operacional e *hardware*. Os parâmetros desse modelo são apresentados na Tabela 7.

Figura 21 – MODELO DO CONTROLADOR NODE

PARÂMETROS	DESCRIÇÃO
MTTFHyper-V, MTTFS.O, MTTFHard	Tempo Médio de Defeito de Hyper-V, S.O, Hard
MTTRHyper-V, MTTRS.O, MTTRHard	Tempo Médio de Reparo de Hyper-V, S.O, Hard

Tabela 7 – PARÂMETROS DE MODELO DO CONTROLADOR NODE

5.2.1.4 MODELO VM-MASTER NODE

O modelo RBD do *VM-MASTER NODE* (Figura 22) representa os componentes da máquina virtual do master node. O modo operacional desse modelo é $OM_{PN} = (MN \wedge HYPER - V \wedge S.O)$, onde o MN, HYPER-V, S.O, HARD representam os componentes do master node, *Hyper-V*, sistema operacional da máquina virtual. Os parâmetros desse modelo são apresentados na Tabela 8.

Figura 22 – MODELO DO VM - MASTERNODE

PARÂMETROS	DESCRIÇÃO
MTTFMN, MTTFHyper-V, MTTFS.O	Tempo Médio de Defeito de MN, Hyper-V, S.O
MTTRMN, MTTRHyper-V, MTTRS.O	Tempo Médio de Reparo de MN, Hyper-V, S.O

Tabela 8 – PARÂMETROS DE MODELO DA VM - MASTER NODE

5.2.1.5 MODELO VM-DATANODE

O modelo *VM-DATANODE* (Figura 23) representa os componentes DATANODE, SISTEMA OPERACIONAL e *HYPER-V* e o modo operacional desse modelo é $OM_{PN} = (DN \wedge S.O \wedge HYPER - V)$, onde o DN, S.O e *HYPER-V* representam os componentes data node, sistema operacional e *Hyper-V* do data node. Os parâmetros desse modelo são apresentados na Tabela 9.

Figura 23 – MODELO DO VM - DATANODE

PARÂMETROS	DESCRIÇÃO
MTTFDN, MTTFS.O, MTTFHyper-V	Tempo Médio de Defeito de DN, S.O, Hyper-V
MTTRDN, MTTRS.O, MTTRHyper-V	Tempo Médio de Reparo de DN, S.O, Hyper-V

Tabela 9 – PARÂMETROS DO MODELO DO VM - DATANODE

6 ESTUDO DE CASO

Este capítulo apresenta três estudos de casos, o Estudo de caso 1 tem o objetivo de avaliar o desempenho do *Hadoop cluster* configurado na infraestrutura de nuvem privada. O Estudo de caso 2 tem o objetivo de aprensentar a disponibilidade desse ambiente utilizando modelos RBD combinados através de uma estratégia de modelagem hierárquica. O estudo de caso 3 apresenta os resultados de performabilidade através do método para avaliação de performabilidade de ambientes *big data* na nuvem privada. Com os estudos de caso podemos mostrar que com a utilização do processador e memória mostramos o consumo de cada recurso oferecido na nuvem privada e a disponibilidade dos componentes dos serviços da infraestrutura de nuvem privada através de modelos RBD.

6.1 ESTUDO DE CASO 1

O Estudo de caso 1 tem o objetivo de avaliar o desempenho do *Hadoop cluster* configurado na infraestrutura de nuvem privada, conforme a metodologia e modelos propostos nos Capítulos 4 e 5.

6.1.1 ENTENDIMENTO E CONFIGURAÇÃO DO AMBIENTE BIG DATA NA NUVEM PRIVADA

Essa seção apresenta o ambiente adotado para avaliação de desempenho de ambientes *big data* na nuvem (Figura 22). O ambiente de nuvem privada é composto por 7 computadores, com sistema operacional sem interface gráfica para reduzir o consumo de recursos e com a plataforma de nuvem *CloudStack* (CLOUDSTACK, 2022). Essas máquinas possuem a configuração conforme apresentado na Tabela 10 considerando a configuração mínima para a instalação do *hadoop cluster* e do *cloudstack*.

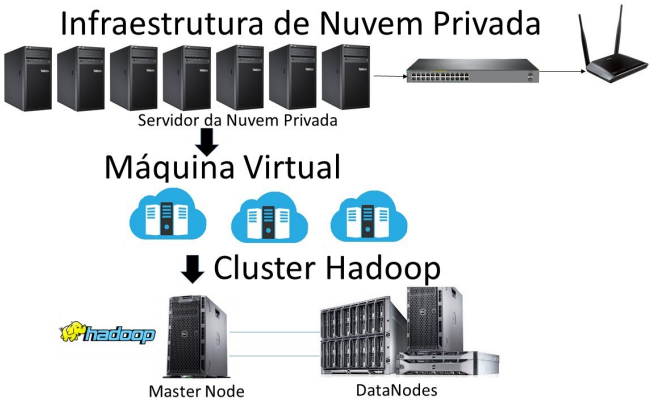

Figura 24 – INFRAESTRUTURA DE NUVEM

Dispositivos	Configuração
Memória	8GB | 4GB
Processador	I5 | I3
HD	1TB
Hypervisor	KVM
S.O	CentoS 7
Plataforma de Nuvem	CloudStack

Tabela 10 – CONFIGURAÇÃO DAS MÁQUINAS QUE COMPÕEM A NUVEM PRIVADA

6.1.2 PLANEJAMENTO DE EXPERIMENTOS DO AMBIENTE BIG DATA NA NUVEM PRIVADA

No planejamento de experimentos foram adotados fatores com diferentes níveis. A oferta de serviço define a capacidade da nuvem privada para os data nodes configurados nas máquinas virtuais. Conforme a Tabela 11 que mostra as ofertas de serviço adotadas para o planejamento de experimentos.

Oferta de Serviço	Configuração
Small	Mem:4GB | Proc: 6 Cores | Armazenamento: 1TB
Medium	Mem:6GB | Proc: 6 Cores | Armazenamento: 1TB
Large	Mem:8GB | Proc: 8 Cores | Armazenamento: 1TB

Tabela 11 – OFERTA DE SERVIÇO DA INFRAESTRUTURA DE NUVEM PRIVADA.

Os níveis selecionados para o planejamento de experimento são mostrados na Tabela 12, foram adotados de acordo com a capacidade de processamento e de memória da nuvem privada.

Cenário	Oferta de serviço	Carga de trabalho	Número de data nodes
1	Small	3GB	3
2	Medium	4GB	4
3	Large	5GB	5

Tabela 12 – CENÁRIO DO PLANEJAMENTO DE EXPERIMENTOS.

6.1.3 GERAÇÃO DE CARGA DE TRABALHO

Para geração da carga de trabalho foi realizada a análise de sentimentos de usuários de redes socias do *Twitter* que fizeram posts usando palavras que pessoas com sintomas de depressão adotam. Entre o período de 05 a 19 de dezembro de 2020 foram coletados 5GB dados, que foram convertidos em um conjunto de dados. A ferramenta utilizada neste processo de coleta de dados foi o *RStudio* (RSTUDIO, 2022) que é uma aplicação que consegue capturar dados do *Twitter* com a execução de um script, através de um pacote chamado twitteR (RSTUDIO, 2022). Nesse ambiente, o map reduce é adotado para analisar um conjunto de dados composto de frases de usuários da rede social *Twitter* que fizeram posts adotando palavras que denotam sintomas de depressão. De acordo com a Organização Pan-americana de Saúde (OPAS, 2022), os transtornos mentais são responsáveis por aproximadamente 13 % das doenças no mundo, em mais de 300 milhões de pessoas, de todas as idades acometidas com esse transtorno. A depressão é uma doença incapacitante funcionalmente e socialmente, que incluem sintomas e períodos diferenciados para cada indivíduo. A captura das publicações desta rede social foi realizada de acordo com parâmetros de análise de sentimentos pesquisados na literatura e no site do ministério da saúde do governo federal do Brasil. As palavras adotadas por pessoas que podem ter sintomas de depressão são: triste, ansioso, depressivo, depressão, esquizofrenia, saúde mental, ansiedade, terapia, doença mental, deprimido, nervoso, incomodado, estressado, envergonhado, arrepender, frustrado, insatisfeito, triste (GOV, 2022). A Figura 23 mostra

o script que foi utilizado para coleta dos dados utilizando a ferramenta proposta o *RSSTUDIO*, onde as linhas 1, 2, 3, 4, 5, 6 e 7 da Figura 23 proporcionaram a instalação e a requisição do pacote necessário para o acesso aos dados da rede social. As linhas 9 até 13 indicam o fornecimento das credenciais *(API KEY, API SECRET, ACCESS TOKEN e ACCESS SECRET)* para o acesso aos dados da rede social e sua autenticação. A linha 14 indica o período de coleta do conjunto de dadoss. A linha 15 indica quais termos devem fazer parte de comentários da rede social, que neste caso são hashtags relacionadas a análise de sentimentos. As linhas 16 a 19 armazenam na variável tweets uma busca de 1000 posts da rede social. A linha 20 tem a função de estruturar os dados de comentários da rede social para processamento no ambiente *Hadoop cluster* (HADOOP, 2022). Por fim, a linha 23 salva o conjunto de dados em um arquivo no HDFS.

```
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”)
```

Figura 25 – SCRIPT UTILIZADO PARA COLETA DOS DADOS COM O TWITTER

6.1.4 MEDIÇÃO DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Nessa atividade, dos experimentos foram replicados 30 vezes, e os resultados das médias das métricas da utilização do processador e utilização de memória foram computados. A Tabela 13 apresenta os resultados das médias das métricas de utilização de processador e utilização de memória dos data nodes configurados em máquinas virtuais

instanciadas na nuvem privada e os resultados das métricas utilização de processador e utilização de memória calculadas através do modelo de desempenho proposto. Nessa tabela, C significa os cenários, UPMed significa a média da utilização do processador medido, UPMod representa a utilização do processador obtida através do modelo, UMMed denota a média da utilização da memória medida, UMMod significa utilização da memória obtida através do modelo, OF representa Oferta de serviço, CG significa a Carga de trabalho e DN representa o número de data nodes através da ferramenta *Minitab*.

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

Tabela 13 – METRICAS DE UTILIZAÇÃO DO PROCESSADOR E MEMORIA

6.1.5 ANÁLISE ESTATÍSTICA DE MÉTRICAS DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Em cada experimento, foram coletadas as métricas utilização de processador e utilização de memória conforme a atividade relacioannda na metodologia. Posteriormente, houve a remoção dos *outliers* das métricas de desempenho com a ferramenta Minitab. Foram calculadas as médias dessas métricas. A análise estatística dos experimentos foi realizada com o auxílio do *software Minitab* (MINITAB, 2022).

6.1.6 MAPEAMENTO DE MÉTRICAS DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

As métricas de desempenho representadas no mapeamento são as utilizações de processador e de memória dos *data nodes* através de tempo de execução configurados em máquinas virtuais da nuvem privada, uma vez que essas métricas proporcionam o planejamento da infraestrutura do *Hadoop cluster* (HADOOP, 2022).

6.1.7 REFINAMENTO DO MODELO DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Para este trabalho o refinamento do modelo de desempenho é realizado através da técnica de aproximação de fases, que calcula o primeiro e segundo momento da distribuição de probabilidade empírica dos tempos de execução do *hadoop cluster*. O refinamento do modelo de desempenho ocorreu com a parametrização deste modelo considerando a distribuição de probabilidade hipoexponencial que representa a métrica tempo de execução coletada em 30 replicações de cada um dos 45 experimentos planejados conforme a Seção 6.1.2 (Tabela 11).

As médias, desvio-padrões e inversos dos coeficientes de variação dos tempos de mapeamento e tempos de redução para os Cenários com 3 data nodes(Tabela 13) foram calculados para identificar a distribuição poliexpolinomial que melhor representa o comportamento desses tempos.

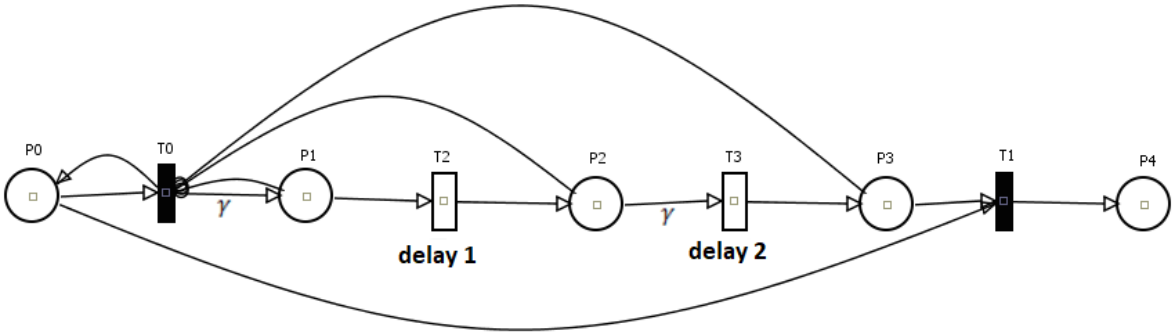

Figura 26 – MODELO DE DISPONIBILIDADE DO HADOOP CLUSTER CONFIGURADO NA NUVEM PRIVADA

A Tabela 14 mostra a distribuição de probabilidade que melhor representa os tempos de execuçõ e de redução do *dataset*, com base no inverso do coeficiente de variação(1/C v) desses tempos. O modelo de desempenho proposto foi refinado com base nos parâmetros da distribuição hipoexponencial. A Tabela 15 apresenta os parâmetros da distribuição Poliexponencial do refinamento do modelo.

Tempos	Tempo médio(segundos)	Desvio-padrão	1/C v	Dist. Prob.
Execução	604,37	287,6	1,25	Hipoexponencial

Tabela 14 – MÉDIA, DESVIO-PADRÃO E DISTRIBUIÇÃO DE PROBABILIDADE EXPOLINOMIAL.

Nº de Fases	Delay 1	Delay 2	Dist. Prob.
4	208.78	98.89	Hipoexponencial

Tabela 15 – PARÂMETROS DA DISTRIBUÍÇÃO POLIEXPOLINOMINAL DO REFINAMENTO DO MODELO.

6.1.8 VALIDAÇÃO DO MODELO DE DESEMPENHO DE AMBIENTES BIG DATA NA NUVEM PRIVADA

A validação do modelo de desempenho proposto foi realizada através do teste T pareado que compara a média da diferença entre duas amostras independentes, nesse caso,

as métricas utilização de processador e utilização de memória medidos nos 45 experimentos e calculados através do modelo de desempenho proposto (GUPTA B. C.; GUTTMAN,). Considerando um nível de significância de 5, o teste t-pareado gerou um intervalo de confiança de (-1,073;1,057) para a métrica utilização de memória e (-3,20;1,45) para a métrica utilização de processador. Como o intervalo de confiança contém 0, não há evidências estatísticas para rejeitar a hipótese de equivalência entre valores medidos e obtidos do modelo de desempenho.

6.1.9 ANÁLISE DE NOVOS CENÁRIOS DE AMBIENTES BIG DATA NA NUVEM PRIVADA

O novo cenário avaliado considera a carga de trabalho máxima suportada pelo *Hadoop cluster* (HADOOP, 2022) com 3, 4 e 5 data nodes configurados na infraestrutura de nuvem privada. Neste estudo, os fatores e níveis são apresentados na Tabela 16.

Soluções	Oferta de serviço	Carga de trabalho	Número de data nodes
1	Large	10GB	3
2	Large	15GB	4
3	Large	25GB	5

Tabela 16 – OFERTA DE SERVIÇO DA INFRAESTRUTURA DE NUVEM PRIVADA

A Tabela 17 apresenta a utilização de processador e utilização de memória dos data nodes configurados nas máquinas virtuais da nuvem privada. Pode-se observar que maiores intensidades de carga de trabalho são aplicadas, considerando o ambiente *big data* com 10, 15 e 25 *data nodes*. Novamente, nessa tabela, C significa os cenários, UPMod representa a utilização de processador obtida através do modelo, UMMod significa a média da utilização da memória obtida através do modelo, OF representa Oferta de serviço, CG significa a Carga de trabalho e DN representa o número de *Data Nodes*.

A carga de trabalho apresenta impacto na variação da métrica utilização de processador e de memória nos Cenários 1 até 9. Nesses Cenários a métrica utilização de processador não atingiu o nível de saturação. Entretanto, a métrica utilização de memória chegou ao nível de saturação em todos os cenários, com valores de utilização maiores que

C	DN	OF	CG	UPMod	UMMod
1	3	LARGE	10GB	72,68	99,11
2	4	LARGE	10GB	72,98	99,36
3	5	LARGE	10GB	77,43	99,55
4	3	LARGE	15GB	72,62	99,03
5	4	LARGE	15GB	76,03	99,43
6	5	LARGE	15GB	77,40	99,51
7	3	LARGE	25GB	72,72	99,16
8	4	LARGE	25GB	75,97	99,35
9	5	LARGE	25GB	72,42	99,11

Tabela 17 – MÉTRICAS DE UTILIZAÇÃO DO PROCESSADOR E MEMÓRIA DOS NOVOS CENÁRIOS ESTUDO DE CASO

99 % (CILENDO, 2005), indicando uma necessidade de redimensionamento desse recurso para evitar perda de desempenho no processamento do *dataset* a configuração utilizada está descrita na Tabela 17.

6.2 ESTUDO DE CASO 2

O Estudo de caso 2 tem o objetivo de avaliar a disponibilidade do *Hadoop cluster* configurado na infraestrutura de nuvem privada, conforme o método para avaliação de disponibilidade de ambientes *big data* na nuvem privada, a estatégia de modelagem hierárquica e os modelos RBD propostos nos Capítulos 4 e 5. As próximas seções apresentarão todas as atividades necessárias para a avaliação da disponibilidade do *Hadoop cluster* configurado na nuvem privada.

6.2.1 ENTENDIMENTO E CONFIGURAÇÃO DO AMBIENTE BIG DATA NA NUVEM PRIVADA

A plataforma *CloudStack* foi configurada com 7 servidores, um destinado ao controlador de nuvem (CLC) e os demais controlador de nó (CN) ver seção 6.1.1 do

Estudo de caso 1. Diferentes máquinas virtuais foram configuradas nos servidores que executam os serviços dos NC's.

6.2.2 GERAÇÃO DE MODELOS DE DISPONIBILIDADE

Nesta seção foram gerados modelos para avaliar a disponibilidade de ambientes *big data* na nuvem privada através da combinação de modelos baseados em RBD conforme a estatégia de modelagem hierárquica proposta (Seção 5.2).

6.2.3 PARAMETRIZAÇÃO DOS MODELOS DE DISPONIBILIDADE

Nesta seção foram parametrizados os modelos RBDs utilizando os valores do MTTFs e MTTRs dos componentes da nuvem privada e do ambiente *big data*. A figura 16 descreve o modelo de disponibilidade do *Hadoop Cluster* configurado na nuvem privada.

O modelo de *Hardware* (Figura 19) representa os componentes memória, processador e disco e é adotado para calcular o MTTF e MTTR do sistema computacional, conforme Tabela 18 (INTEL, 2022). Os valores de MTTF e MTTR calculados para o sistema computacional são 279069 horas e 8 horas.

COMPONENTES	MTTF	MTTR
Processador	1.500.000	8
Memória	480000.0	8
Disco	1.200.000	8

Tabela 18 – VALORES DE MTTF E MTTR DOS COMPONENTES DO HARDWARE

O modelo de Controlador (Figura 20) é composto dos componentes controlador de nuvem e é adotado para calcular o MTTF e MTTR do controlador de nuvem, conforme (Tabela 19). Os valores de MTTF e MTTR calculados para o sistema computacional 16915 horas e 8 horas (MELO, 2017).

O modelo de Controlador de nó (Figura 21) é composto dos componentes monitor de máquina virtual, sistema operacional e sistema computacional e é adotado para calcular o MTTF e MTTR do controlador de nó são, conforme Tabela 20. Esses valores de MTTF

COMPONENTES	MTTF	MTTR
Sistema Operacional CentoS	42000.0	8
Módulo de Gerenciamento	42000.0	8
Hardware	19530.33	8

Tabela 19 – VALORES DE MTTF E MTTR DOS COMPONENTES DO CONTROLADOR

e MTTR calculados para o sistema computacional 2704 horas e 8 horas (MELO, 2017).

COMPONENTES	MTTF	MTTR
Monitor de Máquina Virtual	2990.0	8
Sistema Operacional	42000.0	8
Hardware	19530.33	8

Tabela 20 – VALORES DE MTTF E MTTR DOS COMPONENTES DO CONTROLADOR DE NÓ

O modelo de *VM-Master Node* (Figura 22) é composto dos componentes *master node*, monitor de máquina virtual e sistema operacionale é adotado para calcular o MTTF e MTTR da máquina virtual do *master node*, conforme Tabela 21. Os valores de MTTF e MTTR calculados para a máquina virtual do *master node* são 2064 horas e 8 horas (MELO, 2017).

COMPONENTES	MTTF	MTTR
Master Node	42000.0	8
Monitor de Máquina Virtual	2990.0	8
Sistema Operacional	42000.0	8

Tabela 21 – VALORES DE MTTF E MTTR DOS COMPONENTES DO VM - MASTERNODE

O modelo de *VM Data Node* (Figura 23) é composto dos componentes monitor de máquina virtual, sistema operacional e sistema computacional e é adotado para calcular

o MTTF e MTTR da máquina virtual do *data node*, conforme Tabela 22. Os valores de MTTF e MTTR calculados para a máquina virtual do *data node* são 2767 horas e 8 horas (MELO, 2017).

COMPONENTES	MTTF	MTTR
Data Node	329067.64	8
Sistema Operacional	42000.0	8
Monitor de Máquina Virtual	2990.0	8

Tabela 22 – VALORES DE MTTF E MTTR DOS COMPONENTES DO VM - DATANODE

O modelo de plataforma de nuvem (Figura 18) é composto dos componentes *nobreak*, roteador, *switch*, controlador, controlador de nó, máquina virtual do *masternode* e máquina virtual do *datanode*. Através desses componentes, seus valores de MTTR e TTF são calculado o MTTF e MTTR da disponibilidade da plataforma de nuvem conforme Tabela 10. Os valores de MTTF e MTTR calculados para a plataforma de nuvem são de 2767 horas e 8 horas (MELO, 2017).

COMPONENTES	MTTF	MTTR
NB	329067.64	8
RT	42000.0	8
SW	2990.0	8
CLC	16915.0	8
CN	2704.0	8
MVM	2064.0	8
MVD	2990.0	8

Tabela 23 – VALORES DE MTTF E MTTR DOS COMPONENTES DA PLATAFORMA DE NUVEM

6.2.4 ANÁLISE DE NOVOS CENÁRIOS DE AMBIENTES BIG DATA NA NUVEM PRIVADA

Nesta seção podemos avaliar o impacto na disponibilidade da variação percentual de 50% para mais e 50% para menos do MTTF de cada componente da nuvem privada, conforme a Tabela 24. Mostrando o impacto dos componentes de confiabilidade da nuvem, possuindo uma maior confiabilidade dos recursos com um tempo maior de falha e quando for para -50% o tempo de falha diminui. Da mesma forma a disponibilidade do *Hadoop cluster* na nuvem privada mostrando os valores de disponibilidade sem ajustes no MTTF, com ajuste de +50% no MTTF e com ajuste de -50% no MTTF.
A Tabela 24, mostra os resultados após os ajustes de +50% e -50% de disponibilidade do MTTF.

COMPONENTES	MTTF	MTTR	+50% MTTF	-50% MTTF
NB	329067.64	8	493601.46	164533.5
RT	42000.0	8	63000.0	2100.0
SW	2990.0	8	4485.0	1495.0
CLC	16915.0	8	29295.0	1044.06
CN	2704.0	8	4145.46	1385.18
MVM	2064.0	8	3490.91	1308.67
MVD	2990.0	8	4151.71	1383.90

Tabela 24 – VALORES DE MTTF E MTTR COM VARIAÇÃO PERCENTUAL DE 50% PARA MAIS E 50% PARA MENOS DO MTTF DE CADA COMPONENTE DA PLATAFORMA DE NUVEM PRIVADA.

6.3 ESTUDO DE CASO 3

O objetivo desta seção é apresentar os resultados de performabilidade através do método para avaliação de performabilidade de ambientes *big data* na nuvem privada. Essa avaliação ocorreu através da combinação dos resultados das métricas de desempenho dos cenários da Tabela 13 do Estudo de Caso 1 com o resultado da métrica de disponibilidade apresentada no Estudo de Caso 2. Essa composição provê métricas de performabilidade, as quais foram calculadas, independentemente, a partir do modelo de desempenho e dos modelos de disponibilidade e posteriomente combinadas para mostrar o efeito da disponibilidade no desempenho do *hadoop cluster* na nuvem privada. A Tabela 25, mostra os resultados de desempenho do Estudo de caso 1 (EC1) e do Estudo de Caso 3 (EC3). Após a combinação dos resultados, podemos observar que no estudo de caso 3 o consumo da utilização do processador e memoria foram menores após a combinação dos estudos de caso.

C	UPMod (EC1)	UMMod(EC1)	UPMod (EC3)	UMMod(EC3)	OF	CG	DN
1	72,68	99,11	47,27	82,73	LARGE	10GB	3
2	72,98	99,36	52,24	87,07	LARGE	10GB	4
3	77,43	99,55	56,23	89,98	LARGE	10GB	5
4	72,62	99,03	62,21	82,95	LARGE	15GB	3
5	76,03	99,43	52,32	87,20	LARGE	15GB	4
6	77,40	99,51	56,76	90,82	LARGE	15GB	5
7	72,72	99,16	79,03	98,79	LARGE	25GB	3
8	75,97	99,35	81,12	99,15	LARGE	25GB	4
9	72,42	99,11	82,79	99,39	LARGE	25GB	5

Tabela 25 – RESULTADO DAS MÉTRICAS DE UTILIZAÇÃO DO PROCESSADOR E MEMÓRIA DOS ESTUDOS DE CASO 1 E ESTUDO DE CASO 3

7 CONCLUSÃO

7.1 CONTRIBUIÇÕES

A principal contribuição desse trabalho foi avaliação de desempenho, disponibilidade e performabilidade que contempla apenas a nuvem privada. Os Modelos de desempenho, disponibilidade e performabilidade que foi proposto para avaliar a utilização de recursos da aplicação da nuvem privada.

As contribuições deste trabalho foram a proposição de uma metodologia e um modelo estocástico e modelos combinatoriais para avaliação de desempenho, disponibilidade e performabilidade de ambientes *big data* na nuvem privada.
Existem diversas soluções de plataformas de código aberto para configurações de nuvens privadas, como o *OpenStack* (OPENSTACK, 2022), *OpenNebula* (OPENNEBULA, 2022), (EUCALYPTUS, 2022) e o (CLOUDSTACK, 2022). Este trabalho adotou a plataforma de nuvem *CloudStack* por ser inteiramente *opensource*. O modelo de desempenho proposto foi baseado em redes de Petri estocásticas e proporcionou a avaliação da utilização do processador, memória e tempo de resposta do *Hadoop cluster* na nuvem privada, considerando diferentes ofertas de serviço, carga de trabalho e número de data nodes. Para geração da carga de trabalho foi realizada a análise de sentimentos de posts de usuários do *Twitter* com palavras que indicavam sintomas de depressão.
O estudo de caso baseado na plataforma *CloudStack* e no Hadoop cluster considerou-se data sets de diferentes tamanhos formados a partir da análise de sentimentos de posts de usuários do *Twitter*. Os resultados do modelo de desempenho mostraram que o tamanho do conjunto de dados é o fator de maior influência na utilização dos recursos do aplicativo de *big data* na nuvem privada.
O modelo de disponibilidade proporcionaram a avaliação da disponibilidade de ambientes *big data* na nuvem privada
O estudo proporcionou a estratégia de modelagem para avaliação de disponibilidade de ambientes *big data* na nuvem privada.

7.2 LIMITAÇÕES

As limitações deste trabalho são:

O tamanho dos *datasets* adotados para a geração da carga de trabalho *Big Data*, que foram experimentais devido às restrições relacionadas à capacidade da computacional da nuvem privada;

O máximo de data nodes configurados em máquina virtuais da nuvem privada que puderam ser instanciados foi 7 data nodes, considerando a infraestrutura real.

Este trabalho limitou-se a estudar algumas métricas relacionadas à dependabilidade e performabilidade das IaaS. Outras métricas relacionadas manutenabilidade não foram levadas em consideração. Assim, como adotar um método de análise de sensibilidade para detalhar o impacto das métricas avaliadas;

7.3 TRABALHOS FUTUROS

Como trabalho futuro, pretende-se avaliar o impacto do desempenho, disponibilidade e performabilidade do *Hadoop cluster* configurado na nuvem híbrida.

Pretendemos avaliar o Desempenho do Apache Spark e Apache Flink usando o Benchmark Hibench-master 7 comparando com a plataformas Apache Hadoop.

Proposta de modelos hierárquicos que representam o funcionamento de um serviço de grandes quantidades de *dataset* para avaliar o desempenho e disponibilidade em nuvem privada *CloudStack*.

Propor modelos SPN e RBD para comparar grandes volumes de dados e obter o melhor desempenho e melhor taxa de disponibilidade em comparação a outros tipos de nuvem computacional feito o *OpenStack, CloudStack e Eucalyptus*.

Referências

ARAÚJO, C. J. M. Tomada de decisão multicritério em infraestruturas como serviço em nuvem: Uma abordagem baseada em modelos de dependabilidade, performabilidade e custo. In: . [S.l.: s.n.], 2019.

ASSUNÇÃO, M. D.; CALHEIROS, R. N.; BIANCHI, S.; NETTO, M. A.; BUYYA, R. Big data computing and clouds: Trends and future directions. In: **Journal of Parallel and Distributed Computing, Elsevier**. [S.l.: s.n.], 2015. v. 79, p. 3–15.

BAUSE, F.; KRITZINGER., P. S. Stochastic petri nets - an introduction to the theory. In: **Universit²at Dortmund**. [S.l.: s.n.], 2002.

BERTONCELLO, G. Um estudo sobre a performance de aplicações big data com deep learning. In: **UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL**. [S.l.: s.n.], 2018.

BOLCH, G. e. a. Queueing networks and markov chains: modeling and performance evaluation with computer science applications. In: **Wiley-Interscience**. [S.l.: s.n.], 2006. p. 896.

CENTOS. **CENTOS**. 2022. Disponível em: <https://www.centos.org/>.

CLOUDSTACK. **CLOUDSTACK**. 2022. Disponível em: <https://cloudstack.apache.org/.>

COLUMBUS. **10 Charts That Will Change Your Perspective Of Big Data's Growth.** 2018. Disponível em: <https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth>.

DANTAS, J. R. Planejamento de infraestrutura de nuvens computacionais para serviço de vod streaming considerando desempenho, disponibilidade e custo. In: **In: IEEE, 2016 IEEE International Conference on Big Data**. [S.l.: s.n.], 2018.

DEAN, J.; GHEMAWAT, S. Mapreduce: a flexible data processing tool. In: **Communications of the ACM, ACM**. [S.l.: s.n.], 2010. v. 53, p. 72–77.

_____. Mapreduce: a flexible data processing tool. In: **Communications of the ACM, ACM**. [S.l.: s.n.], 2010. v. 53, p. 72–77.

DEAN J.AND GHEMAWAT, S. Mapreduce: a flexible data processing tool. In: **Communications of the ACM**. [S.l.: s.n.], 2010. v. 53, n. 1, p. 72–77.

ECKROTH, J. Teaching big data with a virtual cluster. In: **In: ACM. Proceedings of the 47th ACM Technical Symposium on Computing Science Education.** [S.l.: s.n.], 2016. p. 175–180.

EUCALYPTUS. **EUCALYPTUS. Amazon Web Services, Eucalyptus Open Source Cloud Computing Infrastructure - An Overview**. 2022. Disponível em: <https://aws.amazon.com.>

FEITELSON, D. G. Workload modeling for computer systems performance evaluation. In: **Cambridge university press**. [S.l.: s.n.], 2015.

FÉ, I. d. S. design methodology: a status in the swedish manufacturing industry. In: **Dissertação (Mestrado) — Universidade Federal de Pernambuco.** [S.l.: s.n.], 2017. v. 19, p. 18–19.

GOV, S. **SAUDE GOV**. 2022. Disponível em: <https://www.gov.br/saude/pt-br>.

GREMYR, I.; M, A.; P., J. design methodology: a status in the swedish manufacturing industry. In: **Quality and Reliability Engineering**. [S.l.: s.n.], 2003. v. 19, p. 285–293.

GUPTA B. C.; GUTTMAN, I. In: **John Wiley Sons**. [S.l.: s.n.].

HADOOP, A. **10 Charts That Will Change Your Perspective Of Big Data's Growth.** 2022. Disponível em: <https://hadoop.apache.org/>.

HASHEM, I. A. T.; YAQOOB, I.; ANUAR, N. B.; MOKHTAR, S.; GANI A.AND KHAN, S. U. The rise of "big data" on cloud computing: Review and open research issues. In: **Information systems, Elsevier**. [S.l.: s.n.], 2015. v. 47, p. 98–115.

_____. The rise of "big data" on cloud computing: Review and open research issues. In: **Information systems, Elsevier**. [S.l.: s.n.], 2015. v. 47, p. 98–115.

INTEL. **INTEL**. 2022. Disponível em: <https://www.intel.com.br/content/www/br/pt/products/details/processors/core/i5/docs.html?s=Newest/>.

JAIN, R. **The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling.** [S.l.]: John Wiley, 1991.

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

KHALIFA, A.; ELTOWEISSY, M. Collaborative autonomic resource management system for mobile cloud computing. In: **IARIA, Proceedings of the Fourth International Conference on Cloud Computing GRIDs, and Virtualization**. [S.l.: s.n.], 2013. p. 115–121.

KIM, D. S.; MACHIDA, F.; TRIVEDI, K. S. Availability modeling and analysis of a virtualized system. In: **PRDC'09. 15TH IEEE PACIFIC RIM INTERNATIONAL SYMPOSIUM ON. Anais..** [S.l.: s.n.], 2009. p. 365–371.

KUO, W.; ZUO, M. J. Optimal reliability modeling: principles and applications. In: **John Wiley Sons.** [S.l.: s.n.], 2003.

LI, Y. X. L.; LIN, Y. A two-stage stochastic model for cloud computing simulation of resources cost-effectiveness analysis. In: **In: IEEE, 2017 3rd IEEE International Conference on Computer and Communications (ICCC)**. [S.l.: s.n.], 2017. v. 19, p. 2267 –2274.

LILJA, D. J. Measuring computer performance: a practitioner's guide. In: **Cambridge university press**. [S.l.: s.n.], 2005.

LIU B.; CHANG, X. H. Z. T. K. R. R. J. Model-based sensitivity analysis of iaas cloud availability. In: **Future Generation Computer Systems.** [S.l.: s.n.], 2018. v. 83, p. 1–13.

MACHADO, F. N. R. **Big Data O Futuro dos Dados e Aplicações.** [S.l.]: Saraiva, 2018.

_____. **Big Data: O Futuro dos Dados e Aplicações.** [S.l.]: Saraiva, 2018.

MAHESHWARI, A. K. Application of big data to smart cities for a sustainable future. In: **Handbook of Engaged Sustainability, Springer**. [S.l.: s.n.], 2018. p. 1–24.

MARINESCU, D. C. Cloud computing: theory and practice. In: **Morgan Kaufmann)**. [S.l.: s.n.], 2017.

MELO, C. e. a. Capacity-oriented availability model for resources estimation on private cloud infrastructure. In: **In: IEEE. Dependable Computing (PRDC), 2017 IEEE 22nd Pacific Rim International Symposium on.)**. [S.l.: s.n.], 2017. p. 255–260.

MENASCE, D. A.; ALMEIDA, V. A.; DOWDY, L. W. **Performance by design: computer capacity planning by example.** [S.l.]: Prentice Hall Professional, 2004.

MINITAB. **MINITAB**. 2022. Disponível em: <https://www.minitab.com/pt-br/>.

MOLLOY, M. K. Performance analysis using stochastic petri nets.computers. In: **EEETransactions on**. [S.l.: s.n.], 1982. p. 913–917.

OLIVEIRA, A. S. Simf: Um framework de injeÇÃo e monitoramento de falhas de nuvens computacionais utilizando spn. In: . [S.l.: s.n.], 2017.

OPAS. **OPAS**. 2022. Disponível em: <https://www.paho.org/pt/brasil/>.

OPENNEBULA. **OPENNEBULA**. 2022. Disponível em: <https://opennebula.io/>.

OPENSTACK. **OPENSTACK**. 2022. Disponível em: <https://www.openstack.org>.

OUSSOUS, A.; BENJELLOUN, F.-Z.; LAHCEN, A. A.; BELFKIH, S. Big data technologies: A survey. In: **Journal of King Saud University-Computer and Information Sciences, Elsevier**. [S.l.: s.n.], 2018. v. 30, n. 4, p. 431–448.

OUTAMAZIRT, A.; ESCHEIKH, M.; AISSANI, D.; BARKAOUI, K.; LEKADIR, O. In: **Performance analysis of the M/G/c/c + r queuing system for cloud computing data centres**. [S.l.: s.n.], 2018.

PULIAFITO A.; RICCOBENE, S. S. M. Evaluation of performability parameters in client-server environments. In: **The Computer Journal**. [S.l.: s.n.], 1996. v. 39, n. 8, p. 647–662.

RAHMAN, H.; BEGUM, S.; AHMED, M. U. Ins and outs of big data: A review. In: **The 3rd EAI International Conference on IoT Technologies for HealthCare**. [S.l.: s.n.], 2016. p. 1–2.

REISIG, J. W. D. Os conceitos de redes de petri. In: **Berlin**. [S.l.: s.n.], 2014.

RELIAWIKI. **Basics of System Reliability Analysis.** 2017. Disponível em: <http://www.reliawiki.org/index.php/Basics_of_System_Reliability_Analysis.>

RSTUDIO. **RSTUDIO**. 2022. Disponível em: <https://www.rstudio.com/>.

RUPE, J. W. Reliability of computer systems and networks fault tolerance, analysis, and design. In: **John Wiley Sons.** [S.l.: s.n.], 2003. v. 35, n. 6, p. 586–587.

SAHNER R. A.; TRIVEDI, K. . P. A. Performance and reliability analysis of computer systems: an example-based approach using the sharpe software package. In: **Kluwer Academic Publishers**. [S.l.: s.n.], 1996.

SCHENFELD, M. C. Uma arquitetura hÍbrida em um ambiente deinternet das coisas. In: **PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL FACULDADE DE INFORMÁTICA.** [S.l.: s.n.], 2017.

SILVA, B. A. A framework for availability, performance and survivability evaluation of disaster tolerant cloud computing systems. In: **Universidade Federal de Pernambuco**. [S.l.: s.n.], 2016.

STATISTA. **Big data analytics market revenue worldwide in 2019 and 2025**. 2022. Disponível em: <https://www.statista.com/statistics/947745/worldwide-total-data-market-revenue/>.

TAMURA, Y.; YAMADA, S. Software reliability analysis considering the fault detection trends for big data on cloud computing. In: **Springer-Verlag Berlin Heidelberg 2015 , Lecture Notes in Electrical Engineering**. [S.l.: s.n.], 2015. v. 349, p. 1021–1030.

TORRES E.; CALLOU, G. A. E. A hierarchical approach for availability and performance analysis of private cloud storage services. In: **Computing**. [S.l.: s.n.], 2018. v. 100, n. 6, p. 621–644.

VASCONCELOS, B. J. R. Modelos para avaliaÇÃo de disponibilidade e cÁlculo de capacidade de um software de compressÃo de vÍdeo distribuÍdo em nuvem openstack. In: . [S.l.: s.n.], 2019.

VEIGA, R. A. J.; EXPóSITO X. C. PARDO, G. L. T. R.; TOURIFIO, J. Performance evaluation of big data frameworks for large-scale data analytics. In: **In: IEEE, 2016 IEEE International Conference on Big Data**. [S.l.: s.n.], 2016.

WANG Y.; KUNG, L. B. T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. technological forecasting and social change. In: **Elsevier**. [S.l.: s.n.], 2018. v. 126, p. 3–13.

WINDOWS. **Start Windows Reliability and Performance Monitor in a specific standalone mode.** 2022. Disponível em: <<https://docs.microsoft.com/en-us/windows-server/administration/windows-commands/perfmon>>.

XIE, M.; DAI, Y. S.; POH, K. L. Computing system reliability: models and analysis. In: **Kluwer Academic/Plenum Publishers.** [S.l.: s.n.], 2004. p. 293.

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