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

RAJEEV RANJAN YADAV

UMA ESTRATÉGIA PARA AVALIAÇÃO DE DESEMPENHO E CUSTO DE AMBIENTES BIG DATA EM INFRAESTRUTURAS DE NUVENS PRIVADAS

RECIFE-PE
2019

RAJEEV RANJAN YADAV

UMA ESTRATÉGIA PARA AVALIAÇÃO DE DESEMPENHO E CUSTO DE AMBIENTES BIG DATA EM INFRAESTRUTURAS DE NUVENS PRIVADAS

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

ORIENTADORA: Erica Teixeira Gomes de Sousa

RECIFE-PE

2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

Y12u Yadav, Rajeev Ranjan
Uma estratégia para avaliação de desempenho e custo de ambientes big data em infraestruturas de nuvens privadas / Rajeev Ranjan Yadav. – 2019.
115 f. : il.

Orientador: Erica Teixeira Gomes de Sousa.
Dissertação (Mestrado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, BR-PE, 2019.
Inclui referências e apêndice(s).

1. Computação em nuvem 2. Big data - Custos 3. Desempenho - Avaliação 4. Petri, Redes de 5. Processo estocástico I. Sousa, Erica Teixeira Gomes de, orient. II. Título

CDD 004

RAJEEV RANJAN YADAV

UMA ESTRATÉGIA PARA AVALIAÇÃO DE DESEMPENHO E CUSTO DE AMBIENTES BIG DATA EM INFRAESTRUTURAS DE NUVENS PRIVADAS

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

Aprovada em:

BANCA EXAMINADORA

Erica Teixeira Gomes de Sousa (Orientadora)
Universidade Federal Rural de Pernambuco
Departamento de Computação - DC

Fernando Antonio Aires Lins
Universidade Federal Rural de Pernambuco
Departamento de Computação - DC

Eduardo Antonio Guimaraes Tavares
Universidade Federal de Pernambuco
Centro de Informática - CIn

Paulo Romero Martins Maciel
Universidade Federal de Pernambuco
Centro de Informática - CIn

À minha família.

À professora Erica Sousa pela orientação.

À Rafaela Renata Silva Nascimento.

Agradecimentos

O meu agradecimento vai para todos aqueles que me apoiaram e me incentivaram neste caminho.

Agradeço à Professora **Érica Sousa** pela orientação e pelos incentivos que contribuíram para esta dissertação. Agradeço também pelo crescimento pessoal e acadêmico proporcionado ao longo da orientação.

Agradeço aos professores **Fernando Lins** e **Gustavo Callou** pela colaboração em artigos aceitos em periódicos.

Agradeço à **Vladimir Silva** pela contribuição em artigos.

Agradeço à **Yogendra Yadava**, meu pai, à **Dayawanti Yadav**, minha mãe, e **Ranjana Yadav**, minha irmã, pelo apoio e suporte.

Agradeço à **FACEPE** por financiar este projeto.

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

"A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo".

(Albert Einstein)

Resumo

A computação em nuvem está emergindo como o principal mecanismo para processar grandes quantidades de dados de forma eficiente. Nesse contexto, as nuvens privadas fornecem infraestruturas que suportam a análise de conjuntos de dados gerados por diferentes fontes, como redes sociais, dados de saúde e dados climatológicos. Compreender como a análise de *Big Data* se comporta em infraestruturas de nuvem privada, é uma abordagem importante para identificar fatores críticos para o desempenho e custo nestes ambientes. Neste contexto, a avaliação de desempenho e custo propicia o suporte para o gerenciamento destes ambientes considerando métricas de desempenho, como tempos de execução e utilização de processador e de memória de máquinas virtuais da nuvem privada, e métricas de custo, como custo de infraestrutura, de consumo de energia elétrica e de *software*. Este trabalho apresenta uma estratégia baseada em uma metodologia e modelos para avaliação de desempenho e de custo de ambientes que executam transações *Big Data* suportados por um *pool* de recursos provisionados pela infraestrutura de nuvem privada. Uma metologia foi proposta para avaliação de desempenho e custo de ambientes big data na nuvem privada. Essa metodolgia contempla atividades como entendimento e configuração do ambiente Big data na nuvem privada, planejamento de experimentos, medição de desempenho e consumo de energia, modelagem de desempenho e modelagem de custo. Um modelo de desempenho baseado em redes de Petri estocásticas é proposto para avaliar a utilização de processadores e memória de máquinas virtuais e os modelos de custo consideram o custo para implantar uma nuvem privada, custos associados ao consumo de energia da análise dos *data sets* e custos de relacionados à aquisição de *software*. O estudo de caso ilustra a aplicabilidade da metodologia, do modelo de desempenho e dos modelos de custo em uma nuvem privada e fornece informações importantes sobre o desempenho e custo, como identificação de fatores que mais impactam na utilização de processadores e de memória de máquinas virtuais e no consumo de energia nestes ambientes. O estudo de caso considerou a análise de um *data set* composto de opiniões de usuários da rede social *Twitter* sobre as eleições presidenciais do Brasil em 2018.

Palavras-chave: Computação em nuvem. *Big Data*. *Hadoop cluster*. Avaliação de desempenho. Avaliação de custo. Redes de Petri estocásticas.

Abstract

Cloud computing is emerging as the main mechanism for processing large amounts of data. In this context, private clouds provide efficient infrastructures that support the analysis of data sets generated by different sources, such as social networks, health data, and climatology data. Understanding how Big Data parsing behaves in private cloud environments is an important approach to identifying critical performance and cost factors in these environments. Performance and cost evaluation provides support to manage these environments by considering performance metrics such as processor and memory utilization of virtual machines, and cost metrics such as infrastructure cost, power consumption cost and software cost of these environments. This paper presents a strategy based on a methodology and models to evaluate Big Data transactions supported by a pool of resources provided by the infrastructure of the private cloud. A methodology is proposed to evaluate the performance and cost of Big Data environments in private clouds. This methodology presents activities such as understanding and configuring the Big Data environment in the private cloud, the design of experiments, performance, and energy consumption measurement, performance modeling and cost modeling. A performance model is based on stochastic Petri nets is proposed to estimate resources utilization of virtual machines and cost models consider the cost of deploying a private cloud, costs associated with the energy consumption of the data set analysis, and costs related to the acquisition of related software. The case study illustrates the applicability of the methodology, performance model, and cost models in a real private cloud and provides important information about these topics, such as identifying factors that most impact processor utilization and virtual machine memory, and the energy consumption in these environments. The case study considered the analysis of a data set composed by opinions of the Twitter social network users regarding the 2018 Brazilian's presidential election.

Keywords: Cloud computing. Big data. Hadoop cluster. Performance evaluation. Performance modeling. Cost evaluation. Cost modeling. Sensitivity analysis. Stochastic Petri nets

Lista de Figuras

Lista de tabelas

Lista de Siglas

BaaS	*Big Data-as-a-service*
CPU	*central processing unit*
DoE	*Design of experiments*
DNA	*deoxyribonucleic acid*
GB	*Gigabytes*
HD	*hard disk*
HDFS	*Hadoop distributed file system*
I/O	taxa de entrada e saída
IaaS	*Infrastructure-as-a-service*
IBM	*international business machines*
imed.	imediata
IoT	*internet of things*
IP	*internet protocol*
IS	*infinite-server*
KVM	*kernel-based virtual machine*
kWh	*kiloWatt* · hora
MC	modelo de custo
MD	modelo de desempenho
NIST	*National institute of standards and technology*
PaaS	*Platform-as-a-service*
PE	planejamento de experimentos
QoS	*quality of service*
RAM	*random access memory*
RdP	rede de Petri
SaaS	*Software-as-a-service*
SLAs	*service level agreements*

SPNs	*Stochastic Petri nets*
SS	*single-server*
T	*transition*
T.I.	Tecnologia da informação
TB	*Terabytes*
temp	temporizada
Tmap	tempo de mapeamento de dados
Tred	tempo de redução de dados
txt	*text*
u.t.	unidades de tempo
VMs	*virtual machines*

Sumário

1 Introdução

Atividades da sociedade, como o surgimento de redes sociais, estão produzindo cada vez mais dados valorados em todo o mundo. A implantação de análise *Big Data* alterou a visão de como organizações acadêmicas e de negócios obtém informações importantes para suas atividades. As principais vantagens obtidas por iniciativas de análise *Big Data* em empresas são a redução de custos, novas vias para inovação e estabelecer uma tomadas de decisão direcionadas por dados (HASHEM et al., 2015; ASSUNÇÃO et al., 2015; MACHADO, 2018; OUSSOUS et al., 2018). Nestes ambientes, 79% dos executivos afirmam que empresas que não se adequarem a trabalhar com grandes quantidades de dados de forma eficiente, perderão posição competitiva e, além disso, tenderão a uma extinção (COLUMBUS, 2018).

No entanto, a análise de *Big Data* ainda é um desafio para ambientes científicos e de negócios devido ao significativo tempo demandado para a análise dos dados, utilização de recursos e grandes infraestruturas computacionais necessárias para analisar essa quantidade de dados (HASHEM et al., 2015; ASSUNÇÃO et al., 2015; MACHADO, 2018; OUSSOUS et al., 2018). As operações de *Big Data* exigem capacidades que excedem os atuais sistemas convencionais de computação e banco de dados. Neste contexto, a computação em nuvem surge como um mecanismo eficiente e escalável para analisar grandes conjuntos de dados e fornecer informações com menores tempos de execução comparado à sistemas convencionais de *clusterização* (MACHADO, 2018). Este paradigma de computação, contribui para a geração de valor de *Big Data*, oferecendo um *pool* de recursos escaláveis e distribuídos para processamentos de grandes volumes e variedades de dados, que são gerados principalmente por dispositivos móveis(*smartphones*), satélites, redes sociais e sensores(IoT principalmente) (YANG et al., 2017).

O monitoramento de recursos em ambientes de computação em nuvem contribui para que aplicações *Big data* sejam analisadas de forma consistente quanto ao seu desempenho e custo considerando diferentes cargas de trabalho, que podem ser caracterizadas por tamanhos de conjuntos de dados ou diferentes tipos de *data sets*. Serviços de computação em nuvem analisando grandes conjuntos de dados são caracterizados como *big-data-as-a-service* (ASSUNÇÃO et al., 2015). Métricas de desempenho como tempos de execução, tempos de resposta e vazão de requisições são cruciais para a fidelidade do usuário

ao utilizar o serviços de computação assim como *overheads* na utilização de recursos computacionais podem incorrer em custos adicionais e degradações de desempenho (JAIN, 1991; MENASCE et al., 2004; LILJA, 2005).

1.1 Motivação

A computação em nuvem vem se apresentando como uma tendência para analisar grandes quantidades de dados de forma eficiente e eficaz (YADAV et al.,), no qual empresas estão cada vez mais adotando plataformas de nuvem para reduzir custos ao mesmo tempo em que utilizam recursos computacionais no modo *pay-per-use* (NIST, 2018). De acordo com o Gartner's group (WANG et al., 2018), plataformas de nuvem para aplicações *Big Data* são a solução preferida de empresas para provisionarem a análise de grandes quantidades de dados com sistemas de armazenamento distribuído.

Muitas organizações de *T.I.*, como a *Yahoo!* e a *IBM*, adotam o Hadoop para oferecer serviços de análise de dados (OUSSOUS et al., 2018) suportados por infraestruturas de nuvem. O processamento de dados e sua análise representam um desafio para analistas e pesquisadores, sobre quais recursos devem ser gerenciados e controlados, considerando metas de qualidade, desempenho. A computação em nuvem e *Big Data* estão atraindo a atenção da comunidade global e a otimização do uso de recursos em aplicativos de análise é um desafio a ser considerado em pesquisas acadêmicas e ambientes de negócios (KAMTEKAR, 2015; ASSUNÇÃO et al., 2015).

De acordo com o site Statista (STATISTA, 2018), em 2015 o *Hadoop* representava uma parcela de 22,22% das operações de *Big Data* realizadas no mundo. A projeção para 2020, é de que 50% destas operações sejam realizadas por *Hadoop clusters*, contribuindo com 50 bilhões de dólares de valor mercado mundial. O valor de mercado consiste em representar a adoção do *Hadoop* para ambientes *Big Data* considerando todas as empresas que adotam *frameworks Big Data* em suas operações. A Figura 1 apresenta os valores de mercado entre o *Hadoop* e seus concorrentes no mercado *Big Data* e demonstra que o *Hadoop* é a principal *framework* para ambientes *Big Data* que está sendo adotado por empresas para análise de grandes conjuntos de dados.

Kamtekar (KAMTEKAR, 2015) discutiu a importância da análise de *Big Data* suportada por infraestruturas em nuvem. Em ambientes de nuvem, os desafios de

desempenho e eficiência podem ser tratados através da análise e modelagem de utilização de recursos de máquinas virtuais instanciadas no ambiente de nuvem. O modelo de desempenho leva à identificação de fatores que podem afetar o desempenho de aplicações *Big Data*, considerando diferentes níveis de intensidades de carga de trabalho, ofertas de serviço e quantidade de máquinas virtuais instanciadas para o *cluster*. Ainda de acordo com este autor (KAMTEKAR, 2015), a modelagem de *Big Data* em ambientes de nuvem ajuda a melhorar o desempenho e aumentar a eficiência de análise conjuntos de dados, uma vez que pode representar diversos aspectos do sistema baseado em dados de experimentos.

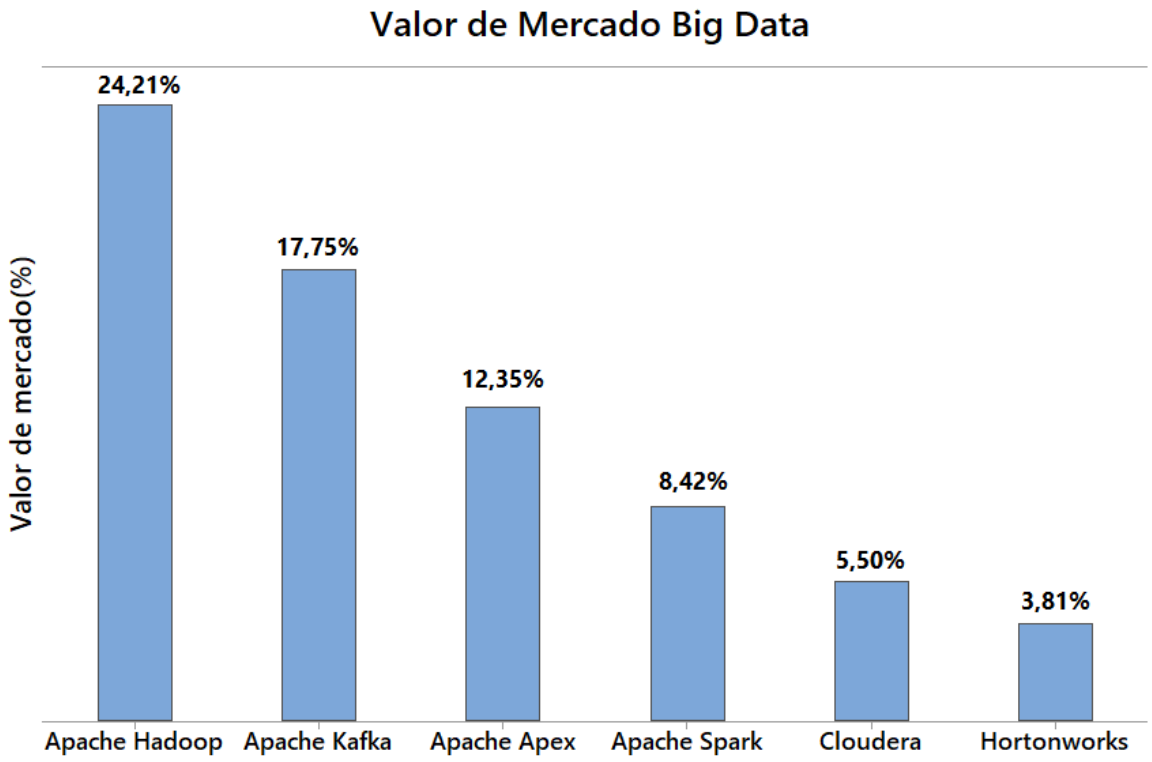

Figura 1 – Valores de mercado do *Hadoop* em relação a seus concorrentes (FONTE: (DATANYZE, 2018))

De acordo com Marinescu (MARINESCU, 2017), as redes de Petri estocásticas podem modelar aplicativos e serviços de computação em nuvem e estimar métricas importantes, como tempos de resposta (YADAV et al.,) e utilização de recursos computacionais da nuvem privada. Modelos de desempenho baseado em redes de Petri estocásticas (SPNs) (MARSAN; BALBO, 1994; MARINESCU, 2017) propiciam a modelagem e avaliam métricas como a utilização de processadores e de memória de máquinas virtuais ao realizarem a análise de grandes quantidades de dados.

O foco deste trabalho é o *IaaS* (*Infrastructure-as-a-Service*) para aplicações *Big Data*, em que um conjunto de recursos computacionais está disponível para os usuários e fornece a infraestrutura necessária para executar a análise de *Big Data*. A interação entre IaaS e transações de *Big Data* é uma abordagem importante para entender melhor o comportamento dos recursos computacionais durante a geração de cargas de trabalho pela análise de grandes quantidades de dados.

1.2 Problema de Pesquisa

Empresas de *Data Analytics* e *Data Analysis* estão se preocupando mais com a arquitetura e implantação de infraestruturas escaláveis, confiáveis e com flexibilidade para analisar dados gerados por diferentes fontes de dados. Com isso, surge a necessidade de efetivo gerenciamento de recursos computacionais para análise de *Big Data* assim como a necessidade da identificação de fatores impactantes para o desempenho e custo destas aplicações (KAMTEKAR, 2015; OUSSOUS et al., 2018).

Overheads de recursos, como processadores e memória, geram degradações de desempenho, resultando em menor fidelização de clientes que solicitam serviços do tipo *Big data-as-a-service-BaaS* (ASSUNÇÃO et al., 2015). A identificação de fatores e a modelagem de desempenho e custo de aplicações *Big Data* implantadas em nuvens privadas auxiliam no entendimento de forma mais clara como ocorre a análise de *Big Data* em ambientes de nuvem privada, considerando aspectos de desempenho e custo. Como estas aplicações consomem recursos de máquinas virtuais que analisam os dados de forma distribuída e qual o custo incorrido destas aplicações são questões fundamentais nestes ambientes.

Para isso, a avaliação e a modelagem de desempenho e de custos baseados em fatores tais como ofertas de serviço da nuvem privada, quantidade de máquinas virtuais na nuvem privada e o tamanho do conjunto de dados(carga de trabalho considerada), podem ser propostas para um melhor gerenciamento da infraestrutura de nuvem privada. O problema de pesquisa que motiva esta dissertação é descrito na seguinte pergunta: “Como avaliar o desempenho e o custo de ambientes *Big Data* configurados em ambientes de nuvens privadas?”.

1.3 Objetivos

O principal objetivo deste trabalho é propor uma estratégia baseada em modelos e uma metodologia para avaliação de desempenho e custo de ambientes *Big Data* em infraestruturas de nuvem privada.

Os objetivos específicos deste trabalho estão descritos a seguir:

- Propor uma metodologia para análise e modelagem de desempenho e custo de aplicações *Big Data* em ambientes de nuvem privada;
- Identificar e hierarquizar fatores impactantes no desempenho e custo de ambientes *Big Data* em nuvem privada;
- Construir um modelo de desempenho para representar ambientes *Big Data* considerando a infraestrutura de nuvem privada;
- Conceber modelos para avaliar o custo de infraestrutura, de software e de consumo de energia de aplicações *Big Data* em ambientes de nuvem privada;
- Gerar cargas de trabalho através da captura e análises de *data sets*
- Conceber um algoritmo para analisar os *data sets* em um ambiente *Big Data*

1.4 Contribuições

A principal contribuição deste propor uma estratégia para avaliação de desempenho e de custo de ambientes *Big Data* implantados em infraestruturas de nuvens privadas.

As demais contribuições desta dissertação são:

- A proposição de um modelo de desempenho e modelos de custos para ambientes *Big Data* suportados por infraestruturas de nuvens privadas;
- Apresentação de fatores que são hierarquizados considerando métricas de desempenho e custo
- Proposição de um modelo de desempenho para cálculo de métricas importantes para estes ambientes são calculadas considerando cenários mais robustos, com maiores intensidades de cargas de trabalho e maiores capacidades de máquinas virtuais instanciadas para analisar os conjuntos de dados na nuvem privada.
- Proposição de modelos de custo contemplando custos para implantar e executar aplicações *Big Data* na nuvem privada como custo de infraestrutura de nuvem privada, custo de consumo de energia elétrica e custo de aquisição de *software*.

- Adoção de técnicas de avaliação e modelagem de desempenho e custo em conjunto com o planejamento de experimentos para representar e estimar aspectos importantes em serviços de *Big Data* (ASSUNÇÃO et al., 2015; IBM, 2018).

1.5 Estrutura do Documento

O presente documento está dividido em 7 capítulos e são descritos brevemente a seguir. O Capítulo 2 apresenta os principais conceitos e terminologias associadas às aplicações *Big Data* que são executadas em ambientes de nuvem privada assim como sua avaliação de desempenho. O Capítulo 3 traz os trabalhos relacionados abordados neste documento e uma comparação demonstrativa das contribuições de cada um. O Capítulo 4 traz a metodologia proposta, com uma sequência definida de atividades necessárias para a avaliação de desempenho e de custo. O Capítulo 5 apresenta os modelos de desempenho e de custo que consideram aspectos importantes para ambientes de análise *Big Data* em nuvens privadas. O Capítulo 6 apresenta o estudo de caso, em que a metodologia proposta é aplicada em um ambiente real de infraestrutura privada suportando cargas de trabalho *Big Data*. Por fim, o capítulo 7 traz as conclusões, contribuições, limitações acerca deste documento e os trabalhos futuros que podem se estender a partir desta dissertação.

2 Fundamentação Teórica

Este capítulo, apresenta os principais conceitos para o entendimento do trabalho proposto. Inicialmente, este capítulo apresenta conceitos sobre computação em nuvem. Esses conceitos estão relacionados aos modelos de implantação e serviços de computação em nuvem. Em seguida, conceitos sobre *Big Data* e avaliação de desempenho são apresentados. Os conceitos sobre avaliação de desempenho contemplam a apresentação de redes de Petri, redes de Petri estocásticas e a técnica de aproximação de fases em redes de Petri estocásticas.

2.1 Computação em Nuvem

A computação em nuvem permite que usuários possam executar aplicações e serviços através da internet se concentrando apenas em suas atividades, deixando a mantenabilidade e gestão de infraestruturas do ambiente para o provedor de nuvem. Sua flexibilidade torna possível que diferentes públicos-alvos sejam clientes dependendo das suas necessidades e interesses, como é o caso dos modelos de implantação classificados como nuvens públicas, privadas, ou híbridas, no qual o cliente decide o nível de controle da infraestrutura de nuvem e seus recursos (ERL et al., 2013).

Nuvens privadas são implantadas geralmente para uma única entidade, que possui todo o controle da infraestrutura de nuvem e de como aplicações e serviços são implantados, além de serem responsáveis pela segurança de todos os dados presentes no sistema de armazenamento da nuvem computacional. Nuvens públicas são provisionadas para o público em geral, em que o objetivo é fornecer recursos computacionais e obter lucro pelas utilizações.

Diferente das nuvens privadas, a nuvem pública é útil para organizações que queiram reduzir custos com aquisição de *hardware* e *software* além de otimizar o tempo, focando exclusivamente em suas atividades. Por fim, um ambiente de nuvem que é composto de nuvens privadas e nuvens públicas são chamados de nuvens híbridas, em que dados e aplicações são compartilhados entre eles de forma a obter vantagens pontuais de cada um (MELL et al., 2011; ERL et al., 2013).

A Figura 2 ilustra um ambiente de nuvem privada com usuários requisitando a infraestrutura de acordo com suas necessidades e objetivos.

Figura 2 – Nuvem privada com máquinas virtuais sendo requisitadas na nuvem privada para atender cargas de trabalho (FONTE: Próprio autor)

Aspectos que demonstram a flexibilidade da computação em nuvem para/com os usuários são as três camadas de arquiteturas como serviço que podem ser oferecidos neste paradigma de computação (MELL et al., 2011; ERL et al., 2013):

- *SaaS*: aplicações e serviços são alcançáveis a partir de vários dispositivos no qual o cliente controla apenas configurações da aplicação, sem a visibilidade de gestão de recursos computacionais.
- *PaaS*: usuários são equivalentes à desenvolvedores, que adotam o ambiente de computação em nuvem para construir e testar suas aplicações sem maiores preocupações com a gestão de recursos que são utilizados por suas aplicações.
- *IaaS*: usuários possuem um maior controle sobre a infraestrutura de nuvem, alocando recursos de acordo com sua necessidade e executando aplicações de propósito geral, incluindo sistemas operacionais. Os usuários neste modelo de serviço, são equivalentes à administradores de sistemas.

Usuários do modelo de serviço *IaaS*, são geralmente administradores de sistemas que são responsáveis pelo gerenciamento de recursos de serviços e aplicações que são

oferecidos suportados pela nuvem(*Web services*). Usuários do *PaaS* possuem plataformas completas para desenvolvimento e testes de *software*, com uma menor preocupação sobre o gerenciamento de recursos das máquinas virtuais. Por fim, usuários do *SaaS* possuem o menor grau de controle dentre os modelos de serviço, em que o usuário tem permissão apenas para alterar configurações básicas da aplicação ou serviço em uso.

Existem diferentes alternativas para o gerenciamento de nuvens privadas como *Cloudstack* e *Openstack*. De acordo com Vogel (VOGEL et al., 2016), a plataforma de nuvem privada *Cloudstack* apresenta desempenho melhor ao *Openstack*, em que cargas de trabalho comuns em ambiente *Big Data*, como *Linpack*, são adotadas para a avaliação e comparação de desempenho. O Apache *Cloudstack* (CLOUDSTACK, 2018) é uma plataforma de nuvem de código aberto que permite implantação, gerenciamento e controle de nuvens privadas. Esse orquestrador de nuvem opera com base em uma pilha de serviços, em sua maioria agentes que garantem o funcionamento de serviços essenciais da nuvem privada.

2.2 *Big data*

Empresas e organizações que não se preparam para obter informações relevantes no seu ambiente de concorrência, tendem a ser extintas no mercado (COLUMBUS, 2018). Quanto maior o acesso da tecnologia para a população, mais informações são geradas e consumidas por organizações para tomadas de decisão. As tomadas de decisão baseadas em análise de dados e estão cada vez mais presentes em ambientes de negócios e ambientes acadêmicos. Tecnologias de processamento e gerenciamento de dados, como o *Hadoop* (APACHE, 2018) e *MapReduce* (DEAN; GHEMAWAT, 2010), permitem geração de informações a partir de um grande conjunto de dados, considerando critérios de qualidade, satisfação de *stakeholders*, utilização de recursos, e custo (MACHADO, 2018).

Big data analytics, como é conhecido na comunidade científica, é responsável por processar estes volumes de dados gerados por diferentes fontes com infraestruturas eficientes e técnicas que tragam consequências positivas para todos os envolvidos no planejamento da análise de dados (WANG et al., 2018).

De acordo com Maheshwari (MAHESHWARI, 2018), operações de *Big data* demandam capacidade que excede os atuais sistemas de armazenamento convencionais e

pode ser caracterizado por quatro Vs:

- Volume, denota a crescente quantidade de dados gerados por diversas fontes ao redor do mundo. A redução de custos e melhorias nas tecnologias de dispositivos de armazenamento, contribuíram para este crescimento da quantidade de dados.
- Velocidade se refere ao espaço cada vez menor de tempo em que conjuntos de dados são gerados. Redes sociais são um exemplo representativo. Outro ponto importante aqui é que o aumento da velocidade de rede e internet pelo mundo também contribuiu para a velocidade que os dados são enviados e recebidos entre dois pontos.
- Variedade representa o fato de que diferentes fontes de dados fornecem as mais variadas formas de dados, tais como dados vindos de sensores de clima e dados de imagens. Cabe ao pesquisador, saber como tratar cada conjunto de acordo com o propósito do ambiente *Big Data*.
- Veracidade consiste na qualidade dos dados que são analisados com o intuito de gerar informações. A qualidade está relacionada à dados sem distorções, no qual razões como erros técnicos e atividades maliciosas podem incorrer em dados sem qualidade.

Modelo de dados e modelos de processamento são a base para o ecossistema *Big data*, em que os modelos de dados fornecem representações lógicas e estruturais para o processamento e gerenciamento de grandes quantidades de dados. Os modelos de processamento possuem um foco nos aspectos físicos envolvidos na análise de *Big data* tais como utilização de processador, memória, disco, e rede de *hosts*.

Dados recentes das atividades da sociedade são gerados principalmente por ciências da terra, IoT e redes sociais, no qual os principais desafios estão relacionados a como armazenar e processar estes dados de forma eficiente, em que recursos computacionais e consumo de energia elétrica devem ser avaliados. Diante deste fato, diversas tecnologias estão sendo abordados e desenvolvidos, tais como *Hadoop* e *MapReduce* (ASSUNÇÃO et al., 2015; HASHEM et al., 2015; DEAN; GHEMAWAT, 2010; OUSSOUS et al., 2018).

O *Hadoop* é um *framework* que processa grandes quantidades de dados através de mecanismos distribuídos como o *MapReduce* (DEAN; GHEMAWAT, 2010) 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 (HASHEM et al., 2015; ALAPATI, 2016). Possui a capacidade de lidar com grandes conjuntos de dados (de *Gigabytes* a *Petabytes*), enquanto mecanismos de tolerância a

falhas lidam com falhas de hardware. O *HDFS* é o sistema de armazenamento distribuído do *Hadoop* e os *data sets* armazena todos os conjuntos de dados e os distribui pelos *data nodes* do *Hadoop*. O *MapReduce* é adotado para dividir em tarefas menores a análise de grandes quantidades de dados, como mapear partes do *data set* e reduzir os dados, gerando a informação para o usuário em tempos reduzidos comparados à sistemas convencionais de armazenamento (ALAPATI, 2016). A Figura 3 mostra a análise de dados gerados por diferentes fontes e seu processamento para a geração de informações acerca do *data set*.

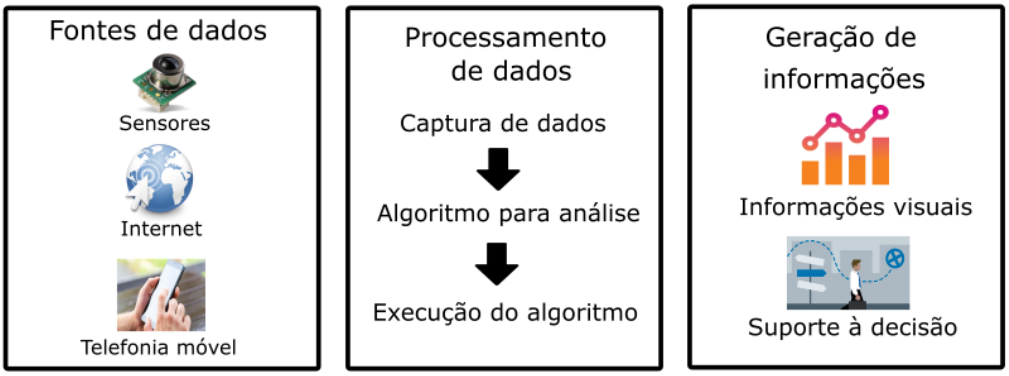

Figura 3 – Análise de dados gerados por diferente fontes (YANG et al., 2017)

No primeiro momento, os dados gerados por diversas fontes, como sensores e redes sociais. Em seguida, estes dados são processados suportados por infraestruturas computacionais e técnicas de análise de dados. *Hadoop*, *MapReduce*, e *Spark* representam as principais *frameworks* para análise de dados de forma distribuída. Por fim, informações de valor, também chamadas de *insights*, são fornecidas e visualizadas. Visualização de análise de grandes conjuntos de dados permite que todas as partes envolvidas possam ter um melhor entendimento dos resultados providos e também contribuem para melhores tomadas de decisão.

2.3 Avaliação de Desempenho

Desempenho é muitas vezes relacionado apenas à velocidade com que serviços e aplicações completam requisições geradas por usuários (JAIN, 1991). Além disso, o desempenho é caracterizado por objetivos multi-critérios em que custo e utilização de recursos também são contemplados. Métricas como vazão e tempos de resposta denotam a qualidade do serviço oferecido pelo sistema (YADAV et al.,) enquanto métricas como

utilização de recursos, utilização de equipamentos e confiabilidade dos componentes da infraestrutura, representam caracteerísticas de gerenciamento (MENASCE et al., 2004; FEITELSON, 2015).

Neste contexto, é necessário que o desempenho seja medido e representado através de forma concreta e visível os resultados para todos os envolvidos(*stakeholders*). Para isso, métricas de desempenho devem ser adotadas para esta representação. (LILJA, 2005). Métricas de desempenho como tempos de execução e utilização de recursos de cargas de trabalho devem ser escolhidas de acordo com os objetivos da avaliação de desempenho. Estes objetivos podem ser compostos da identificação de gargalos sistema ou planejamento de novos sistemas.

Em avaliação de desempenho de sistemas, três técnicas são bastante abordadas (JAIN, 1991; MENASCE et al., 2004; LILJA, 2005; FEITELSON, 2015). A primeira técnica é a base para toda avaliação de desempenho. 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, 2018) 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* (MARSAN; BALBO, 1994). 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.

2.3.1 Redes de Petri

Desenvolvido inicialmente por Carl Adam Petri, redes de Petri podem ser caracterizadas como sendo uma família de elementos gráficos biparticionados adotada para avaliar o comportamento de diferentes sistemas, como uma aplicação sendo executada em ambiente de nuvem. Esta técnica modelagem representa e avalia diferentes aspectos do relacionamento entre componentes de um sistema, incluindo mecanismos de concorrência, sincronização e comunicação. Lugares, representados por círculos, denotam estados locais

que o sistema pode apresentar durante seus diferentes estágios do ciclo de vida. As redes de Petri representam naturalmente modelos probabilísticos e determinísticos (MARINESCU, 2017).

Uma definição mais formal de redes de Petri é de que esta técnica de modelagem consiste em uma tupla formada por um conjunto de lugares, transições, arcos direcionados, pesos destes arcos e estados diferentes que o sistema pode apresentar (MARSAN; BALBO, 1994). São adotadas para avaliar o desempenho de aplicações executadas em ambientes de nuvem e calcular métricas com precisão acerca do ambiente avaliado como tempos de resposta e execução, além de utilização de recursos. Adicionalmente, redes de Petri ainda são recomendadas para modelar sistemas distribuídos e modelar mecanismos de filas que são formadas por requisições de usuários. Possuem diversos tamanhos e formas, sendo compostas de formalismos matemáticos aliados a recursos gráficos (MARINESCU, 2017).

Os elementos que compõem uma rede de Petri são: lugares, denotados por círculos; transições, que representam ações no sistema, e arcos, que denotam o fluxo de marcações do modelo. As marcações dos lugares no modelo *SPN* indicam os diferentes estados que o sistema pode se encontrar em algum instante de avaliação, como avaliar métricas do modeo em estados estacionários (MARSAN; BALBO, 1994). O arco tem a função de estabelecer a dinâmica do fluxo de marcações no modelo. A Figura 4 apresenta os elementos gráficos de uma rede de Petri.

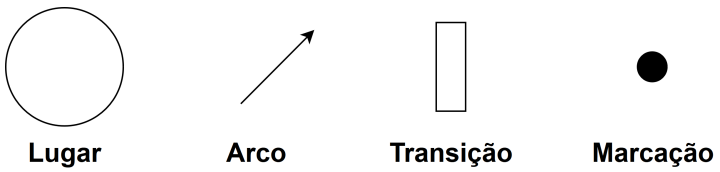

Figura 4 – Elementos de uma rede de Petri (FONTE: Próprio autor)

Uma rede de Petri é caracterizada matematicamente pela tupla $RdP = \{P, T, I, O, M_0\}$ (MARSAN; BALBO, 1994), no qual:

- P denota o conjunto de lugares da rede, que representam componentes e recursos do sistema $P = (p_1, p_2, p_3, ..., p_m)$;
- T representa o conjunto de transições do sistema $T = (t_1, t_2, t_3, ..., t_n)$;
- I é o conjunto de arcos direcionados $I \subset T \times T$;
- O representa o conjunto de funções de peso dos arcos $O \subset T \times P$;
- M_0 representa o estado inicial do sistema $M_0 = (m_{01}, m_{02}, m_{03}, ..., m_{0m})$;

As redes de Petri podem ser analisadas quanto as suas propriedades. As propriedades podem ser comportamentais, que são dependentes da marcação inicial da rede de Petri, e podem ser estruturais, que não dependem da marcação incial da rede (MURATA, 1989).

Dentre as propriedades comportamentais, tem-se:

- Alcançabilidade - indica a chance de uma determinada marcação após um número finito de disparos de transições a partir de uma marcação inicial. Esta propriedade é satisfeita se uma marcação M' é acessível ao existir uma sequência de transições que disparadas, levam a marcação;
- Segurança - é uma particularidade da propriedade de limitação. Se um lugar da rede é k-limitado, significa que o número de marcações acumuladas neste lugar é k;
- Vivacidade - esta propriedade avalia se independente das marcações alcancáveis, seja possível disparar uma transição após uma sequência finita de disparos de outras transições da rede;

Por outro lado, as propriedades estruturais de uma rede de Petri são:

- Estruturalmente limitada - avalia se a rede for limitada para qualquer marcação incial M_0;
- Consistência - avalia se a rede retorna à marcação inicial M_0 após o disparo de uma sequência de transições;

2.3.2 Redes de Petri Estocásticas

Redes de Petri estocásticas(SPNs) são a extensão das redes de Petri(PN) com tempos associados ao modelo, que seguem uma distribuição exponencial. SPNs são compostas de transições temporizadas e imediatas. As transições temporizadas representam aspectos temporizados de sistemas, como uma máquina executando uma tarefa em um determinado período de tempo. Por outro lado, transições imediatas consideram que não há tempo associado à ação que a transição representa, onde o sistema passa de um estado para outro imediatamente (MURATA, 1989; MARINESCU, 2017). Uma rede de Petri estocástica é definida pela 9-tupla $SPN=\{P,\ T,\ I,\ O,\ H,\ \prod,\ G,\ M_0\}$, em que:

- P=$\{p_1, p_2, ..., p_n\}$ é o conjunto de lugares da SPN;
- T =$\{t_1, t_2, ..., t_n\}$ é o conjunto de transições;
- $I \in (\mathbb{N}^n \rightarrow \mathbb{N})^{n \times m}$ é a matriz que corresponde aos arcos de entrada;

- $O \in (\mathbb{N}^n \rightarrow \mathbb{N})^{n \times m}$ representa a matriz que indica os arcos de saída da *SPN*;
- $H \in (\mathbb{N}^n \rightarrow \mathbb{N})^{n \times m}$ denota a matriz relativa aos arcos inibidores;
- $\prod \in \mathbb{N}^m$ é o vetor que associal o nível de prioridades das transições da *SPN*;
- $G \in (\mathbb{N}^n \rightarrow \{true, false\})^m$ representa o vetor associado às condições de guarda relacionada a marcação do lugar à cada transição;
- $M_0 \in \mathbb{N}^n$ é o vetor que representa o estado inicial da *SPN*.

Esta extensão permite que o modelo seja traduzido para uma cadeia de *Markov* de tempo contínuo(CTMC) (MARSAN; BALBO, 1994), no qual métricas podem ser avaliadas através de probabilidades e parâmetros. SPNs permitem o cálculo de probabilidade estacionárias de todas as marcações alcancáveis no modelo e com isso calcular e prever utilização de recursos e tempos associados ao sistema (MARSAN; BALBO, 1994; MACIEL et al., 2017; MARINESCU, 2017).

2.3.3 Técnica 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. 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* (YEE; VENTURA, 2000). O inverso do coeciente de variação dos 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* (JAIN, 1991).

Para identificar qual distribuição expolinomial representa as métricas de desempenho avaliadas, o inverso do coeficiente de variação ($1/C_V$) é adotado(Equação 2.1), no qual μ_d denota a média dos tempos que podem ter seu comportamento estocástico aproximado através da técnica de aproximação de fases enquanto σ_d representa o desvio-padrão destes tempos (YEE; VENTURA, 2000):

$$\frac{1}{C_v} = \frac{\mu_d}{\sigma_d} \tag{2.1}$$

Dependendo do valor de inverso do coeciente de variação $(1/C_v)$ dos tempos medidos, a respectiva atividade tem uma dessas distribuições atribuídas: Erlang, Hipoexponencial ou Hiperexponencial (MACIEL et al., 2017):

1. Quando $1/C_V$ é um número inteiro e diferente de um, os dados empíricos são aproximados como uma distribuição Erlang, onde o comprimento (γ) é calculado pela Equação 2.2.

$$\gamma = (\frac{\mu}{\sigma})^2 \tag{2.2}$$

A nova taxa da transição é calculada pela Equação 2.3:

$$\lambda = \frac{\gamma}{\mu} \tag{2.3}$$

A Figura 5 mostra o modelo de redes de Petri estocástica onde a atividade temporizada tem o comportamento denido por uma distribuição de probabilidade *Erlang* (MACIEL et al., 2017).

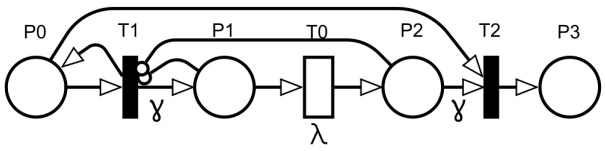

Figura 5 – Rede de Petri estocástica representando uma atividade temporizada com comportamento definido por uma distribuição de probabilidade *Erlang*. Inspirado em (MACIEL et al., 2017)

2. Quando $1/C_V$ é um número maior que um, mas não um inteiro, os dados empíricos devem ser caracterizados como distribuição Hipoexponencial, onde uma SPN é composta de uma sequência cujo comprimento (γ) é calculado por(Equação 2.4):

$$(\frac{\mu_1}{\sigma})^2 \leq \gamma < (\frac{\mu_2}{\sigma})^2 \tag{2.4}$$

E as taxas de transição são calculadas pelas Equações 2.5 e 2.6:

$$\lambda_1 = (\frac{1}{\mu_1}) \tag{2.5}$$

$$\lambda_2 = (\frac{1}{\mu_2}) \tag{2.6}$$

Os tempos médios esperados, μ_1 e μ_2, são calculados pelas Equações 2.7 e 2.8:

$$\mu_1 = \mu \pm \frac{\sqrt{\gamma(\gamma+1)\sigma^2 - \gamma\mu^2}}{\gamma+1} \tag{2.7}$$

$$\mu_2 = \gamma\mu \pm \frac{\sqrt{\gamma(\gamma+1)\sigma^2 - \gamma\mu^2}}{\gamma+1} \tag{2.8}$$

A Figura 6 representa o modelo de redes de Petri estocástica onde a atividade temporizada tem o comportamento denido por uma distribuição de probabilidade Hipoexponencial (MACIEL et al., 2017).

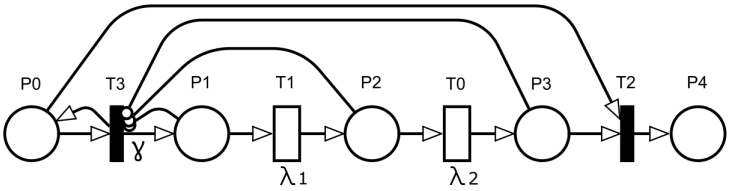

Figura 6 – Rede de Petri estocástica representando uma atividade temporizada com comportamento definido por uma distribuição de probabilidade *Hipoexponencial*. Inspirado em (MACIEL et al., 2017)

3. Quando $1/C_V$ é um número menor que um, a distribuição deve assumir a distribuição Hiperexponencial. Similarmente à rede de distribuição Erlang e Hipo-exponencial, a taxa de transição deve ser calculada pela Equação 2.9, enquanto os pesos (ω_1 e ω_2) das transições imediatas devem ser calculados pelas Equações 2.10 e 2.11:

$$\lambda = \frac{2\mu^2}{\mu^2 + \sigma^2} \tag{2.9}$$

$$\omega_1 = \frac{2\mu^2}{\mu^2 + \sigma^2} \tag{2.10}$$

$$\omega_2 = 1 - \omega_1 \tag{2.11}$$

A Figura 7 representa o modelo de redes de Petri estocástica onde a atividade temporizada tem o comportamento denido por uma distribuição de probabilidade Hiperexponencial (MACIEL et al., 2017).

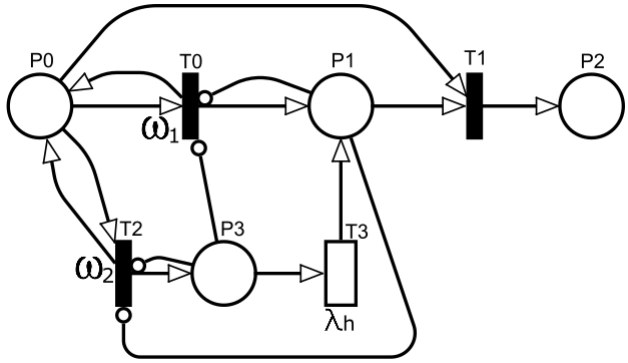

Figura 7 – Rede de Petri estocástica representando uma atividade temporizada com comportamento definido por uma Hiperexponencial. Inspirado em (MACIEL et al., 2017)

2.3.4 Planejamento de Experimentos

Na avaliação de desempenho, os sistemas podem ser abordados como um conjunto de entradas que geram saídas considerando fatores que podem perturbar variáveis de saídas ou também chamadas, variáveis de resposta.

Montgomery (MONTGOMERY, 2017) discute a importância dos experimentos para a ciência e tecnologia de forma geral. Experimentos sempre foram partes importantes de avaliação de desempenho e processos de sistemas, no qual processos são abordados como combinações de operações, poder computacional, máquinas, métodos, e recursos humanos. Na avaliação de desempenho, estas variáveis de resposta são representadas pela utilização de recursos(processador, memória, disco, e energia) e custos associados para suportar diferentes níveis de intensidade de cargas de trabalho enviadas por usuários.

Para Barton (BARTON, 2012), o planejamento de experimentos exerce um importante papel no avanço de ciências de forma geral, em que hipóteses e teorias são colocadas em evidência considerando diferentes ambientes para avaliar o desempenho de sistemas. Os experimentos consideram que variáveis de resposta (Y) possuem valores dependentes de variáveis independentes ou fatores do sistema ($x_0, x_1, x_2, x_3, ..., x_n$), no qual β_0 denota o valor médio dos fatores que não são controlados; β_1, β_2 e β_3 são constantes que são colocadas em prova para determinar o nível de significância de cada fator nas variáveis de resposta; e ε representa erros inerentes aos experimentos. Estes erros geralmente são pequenos com um bom planejamento de experimentos, mas devem ser minimizados ou reduzidos por pesquisadores e administradores de sistemas ao avaliaram o desempenho de sistemas.

As entradas representam cargas de trabalho que o sistema deve executar, como transações geradas por usuários requisitando uma nuvem privada para análise de grandes conjuntos de dados. Os fatores do ambiente são entidades do problema que tem a capacidade de impactar o desempenho do sistema. Este desempenho é representado por multi-critérios que caracterizam o que é de fato, eficiência e custo no sistema a ser avaliado.

Cada fator do sistema abordado para avaliação de seu desempenho possui níveis que são contemplados durante o planejamento de experimentos. Os níveis representam o número de variações que cada fator irá variar durante os experimentos. Com o aumento de níveis dos fatores no problema, a complexidade do planejamento de experimentos também aumenta, em que mais experimentos são necessários para contemplar todos os níveis.

A abordagem correta de acordo com Montgomery (MONTGOMERY, 2017) para avaliar o impacto de dois ou mais fatores em variáveis de resposta é conduzir um planejamento de experimentos **fatorial**. É uma estratégia para avaliar o impacto de fatores em conjunto ao invés de avaliar o impacto de um fator por vez. Esta abordagem propicia o projeto de experimentos, que define quantos experimentos serão executados de acordo com o número de fatores e seus níveis. Outra vantagem desta abordagem é a aleatorização das configurações dos experimentos. O planejamento de experimentos fatorial podem ser completos com cada fator apresentando dois níveis(2^k) ou podem ser fatoriais completos, em que n fatores podem apresentar mais de dois níveis. Adicionalmente, o planejamento de experimentos pode ser fracionado, em que subconjuntos do problema geral são avaliados em instantes diferentes, sendo indicados para ambientes com restrições

de recursos (MONTGOMERY, 2017). O planejamento de experimentos fatorial geral é a abordagem adotada neste trabalho. A Figura 8 denota a interação entre os elementos de um planejamento de experimentos, como os fatores, entradas e saídas(variáveis de resposta) do sistema.

Figura 8 – Planejamento de experimentos considerando entradas, saídas e os fatores do ambiente (FONTE: Próprio autor)

Como resultados importantes do planejamento de experimentos, fatores podem ser hierarquizados de acordo com o nível de impacto considerando as variáveis de resposta assim como predizer variáveis de resposta considerando diferentes configurações do ambiente (MONTGOMERY, 2017).

2.4 Considerações Finais

Este capítulo apresentou os principais conceitos necessários para o entendimento deste documento, abordando avaliação e modelagem de desempenho, planejamento de experimentos, computação em nuvem e *Big Data*. O próximo capítulo apresenta os trabalhos relacionados à esta dissertação.

3 Trabalhos Relacionados

Esta seção apresenta os trabalhos relacionados a esta dissertação. O objetivo é avaliar os métodos e os resultados de estudos que abordam o desempenho e custo de aplicações *Big Data* suportadas por nuvens privadas. Em cada engenho de busca, a seguinte *string* de busca foi adotada: ("*private cloud*" AND "*big data*" AND ("*performance evaluation*"OR "*performance analysis*)" AND ("cost evaluation" OR "*cost analysis*")) com o próposito de avaliar estudos que contemplam estas quatro áreas de estudo. Em razão de obter os estudos mais relevantes em cada engenho de busca, a ordem de relevância foi priorizada e os seguintes critérios de inclusão foram considerados:

- Os artigos devem estar escritos em inglês;
- Os artigos devem ser de periódicos ou conferências;
- Os artigos devem ter sido publicados a partir de 2014.

Após a pesquisa sistemática nas bases científicas *IEEE*, *ACM*, *Springer* e *Elsevier*, os seguintes trabalhos foram selecionados, considerando os critérios de inclusão apresentados assim como o alinhamento entre o trabalho relacionado e a proposta desta dissertação. A seguir, os trabalhos relacionados são apresentados de acordo o tipo de avaliação realizada.

3.1 Avaliação de Desempenho

A avaliação de desempenho de ambientes *Big Data* suportados por infraestruturas de nuvens privadas é importante para a identificação de recursos que podem apresentar *overheads*, gerando gargalos no ambiente. Os trabalhos relacionados desta seção, analisam fatores como ofertas de serviço, quantidade de máquinas virtuais instanciadas para análise de *data sets* e tamanho do *data set* a ser processado no ambiente de nuvem. Cargas de trabalho comuns em ambiente *Big Data* foram geradas e aplicadas, como *Wordcount*, *Terasort* e *Grep*.

Uma abordagem baseada em redes de Petri para avaliar o desempenho e custo de ambientes *Big Data* e *MapReduce* é proposta por Ruiz et al. (RUIZ et al., 2015), no qual análises de *trade-offs* avaliaram fatores de impacto no desempenho como o número de *data nodes*, tempos de processamento e custo de recursos. Com o uso de redes de Petri coloridas, uma extensão de redes de Petri estocásticas, modelos foram gerados para representar o

comportamento destes ambientes considerando seus espaços de estados. Os experimentos foram realizados com a geração de carga sendo caracterizada por um *dataset* de uma rede social. Uma aplicação baseada em *Hadoop* foi responsável pela análise do conjunto de dados em um ambiente de nuvem privada. Os resultados deste trabalho demonstraram que os modelos propostos permitem determinar comportamentos de ambientes *Big Data* suportadas por infraestruturas de nuvens privadas. A carga de trabalho considerada foi de 10 milhões de *tweets*, correspondestes à 1,5 *GB* de dados de uma rede social. O número de *data nodes* apresentou um impacto significativo no tempo de execução da aplicação *Big Data* reduzindo em quase 1 hora os tempos de execução. ao aumentar de um para nove o número de *data nodes* no *Hadoop cluster*, representando uma melhora no desempenho de 9,51% devido a este fator. Resultados do estudo demonstraram que o fator relativo ao número de *data nodes* no *Hadoop* para a análise *Big Data* não apresentou um impacto signitificativo no custo da aplicação *Big Data* na nuvem privada.

Zhang e Sakr (ZHANG; SAKR, 2014) avaliaram o desempenho da análise *Big Data* baseados em *Hadoop* considerando fatores como capacidade de processamento e de memória de *data nodes* assim como o tamanho do *cluster* para analisar de forma distribuída o *dataset*, carcaterizado pela variável de número de *data nodes*. Os cenários são caraceterizados por diferentes níveis de *cores* de CPU, Gigabytes de memória RAM e tamanhos de *clusters* no ambiente *Big Data*. Os níveis adotados para a CPU foram de 1,2 e 4 cores, enquanto para a memória foi de 1 GB, 2 GB e 4 GB. O tamanho do *cluster* ou número de *data nodes* variou entre 1, 2 e 4 *datanodes*. As cargas de trabalho foram geradas por operações comuns em ambientes de análise *Big Data* como *Grep, WordCount e Sort*. O trabalho demonstrou que a alocação de capacidade de máquinas virtuais para ambientes que adotam mecanismos *MapReduce*, possui um importante papel no desempenho e custo destas aplicações. Outra conclusão é de que aplicações *MapReduce* apresentam comportamentos diferentes dependendo se a mesma é *map-intensive* ou *reduce-intensive*. Ainda de acordo com este trabalho, quando há uma intensidade maior de mapeamento de dados nas aplicações do que reduções de *datasets*, maior é o paralelismo associado à análise do conjunto de dados. Este trabalho apresentou resultados de experimentos considerando diversos *benchmarks* para a geração de cargas de trabalho.

Massobrio et al. (MASSOBRIO et al., 2018) apresentaram uma proposta de análise de grandes conjuntos de dados relacionados ao conceito de cidades inteligentes(*smart*

cities) adotando mecanismos de *MapReduce* e *Hadoop* para analisar de forma paralela estes dados em infraestruturas de nuvem. A carga de trabalho foi gerada a partir de dados de transporte público, em que as localizações de cada ônibus são reportadas a cada entre 10 e 30 segundos. Com estes dados, informações importantes para o transporte público foram gerados da análise *Big Data* na nuvem privada. Os níveis de intensidade de carga de trabalho foram 10, 20, 30 e 60 GB representando diferentes meses. Avaliando métricas de desempenho e eficiência como tempos de resposta e consumo de recursos, os resultados deste trabalho mostraram que a abordagem distribuída para análise de dados em ambientes de nuvem pode reduzir os tempos de execução de aplicações de 6 horas para 14 minutos para altas intensidades de carga de trabalho. Este estudo abordou a qualidade do serviço prestado aos usuários do transporte público, que podem ter mais informações em tempo real da localização de ônibus gerados por mecanismos *MapReduce* e *Hadoop*.

Kim e Huh (KIM; HUH, 2017) proporam um sistema chamado *BigPros* para análise de *big data* para o processamento de dados de saúde, com uma metodologia que contempla nuvens híbridas aplicadas em um ambiente de hospital, para análise de dados relacionados à saúde(*healthcare data*) para propriciar um melhor gerenciamento neste ambiente. Esta ferramenta proposta provisionou uma análise eficiente de dados médicos considerando não apenas nuvens privadas, mas também nuvens públicas para suportar transações *big data*. O *cluster Hadoop* possuiu em sua arquitetura 5 máquinas virtuais homogêneas no qual 1 máquina virtual é o *Hadoop master node*, 3 máquinas virtuais são os *Hadoop data nodes* e uma última máquina virtual é um nó mensageiro. A carga de trabalho é representada pelo tamanho do *dataset*, caracterizado através da quantidade de linhas que o conjunto de dados possui(14.000, 35.000, e 70.000 linhas respectivamente). Estes são avaliados de acordo com diferentes níveis de *Split size*, que representa o número de linhas do *dataset* que são processadas a cada iteração do *mapping*. As variáveis de resposta contempladas são o tempo de processamento que o *Hadoop* demandou para terminar a análise dos dados de saúde e o *delay time*, que representa o tempo que um cliente espera na fila. Como resultado das simulações, experimentos foram realizados com o objetivo de avaliar fatores que podem influenciar o desempenho e qualidade de transações neste ambiente. Fatores como a quantidade de memória disponível e a taxa de escrita e leitura do disco da máquina virtual foram identificados como principais fatores que causam atrasos em serviços de *big data* implantados em nuvens.

Adhikari et al. (ADHIKARI et al., 2017) avaliaram o desempenho de *Hadoop clusters* implantandos em nuvem privada e compararam com o desempenho de *Hadoop clusters* implantados diretamente em computadores *commodities*(nativo), sem virtualização. O *Hadoop* foi configurado nas máquinas virtuais na nuvem privada e receberam cargas de trabalho caraterizadas por conversão de imagens em arquivos (PDF). Foi constatado que os computadores *commodities* apresentaram um maior poder computacional e tolerância à falha comparados ao *Hadoop cluster* implantado na nuvem. Porém, a nuvem privada apresentou a vantagem de pode ter escalabilidade e flexibilidade em transações de *big data*, pois, o número de máquinas virtuais que poder ser instanciadas para suportas as cargas de trabalho são escaláveis. As aplicações contempladas para a avaliação de desempenho foram o *WordCount* e a aplicação *Imagetopdf*. O *Hadoop cluster* foi composto de um *master node* e de três *data nodes* com as mesmas capacidades computacionais. A métrica considerada para a avaliação de desempenho foi o tempo demandado para a execução de cada *Hadoop job* para 2 e 4 data nodes. Devido à ausência de virtualização, no ambiente de computadores *commodities*, os tempos de execução foram menores em relação ao ambiente de nuvem privada(cerca de um minuto de diferença). Os autores concluíram que o *Hadoop cluster* suportado por computadores *commodities* apresentou um melhor desempenho considerando as métricas de tempos de execução, mas que, por outro lado, este ambiente não apresentou escalabilidade para suportar diferentes níveis de carga de trabalho comparado ao ambiente de nuvem avaliado. Outra vantagem que o ambiente de nuvem apresentou foram os tempos de *bootup* das máquinas virtuais, que foram menores que os tempos para iniciar máquinas físicas, denotando a flexibilidade da nuvem privada ao gerenciar máquinas virtuais.

Vellaipandiyan e Srikrishnan (VELLAIPANDIYAN; SRIKRISHNAN, 2014) realizaram experimentos com *Hadoop clusters* implantados em ambiente de nuvem. Métricas como *I/O* na rede, CPU, e tempo para completar um *Hadoop job* foram capturados para avaliar o desempenho. Os *datasets* foram gerados a partir de *logs* contendo endereços *IP* de clientes, nomes de usuário, localizações e dentre outros. O *Hadoop cluster* implantado foi composto de 4 *data nodes*, que provisionaram a capacidade computacional para a análise dos *datasets*. As pricipais conclusões deste estudo foram: (1) em *Hadoop clusters*, a memória consumida descresce gradualmente com a variação de parâmetros internos, tais como *block size*(tamanho do bloco) e *split size*(tamanho das partições do *dataset*) analisados. (2) o

aumento do tamanho dos blocos do *dataset*, apresentou impacto positivo nas utilizações de processador e memória de máquinas virtuais na nuvem privada.

Feminella et al. (FEMMINELLA et al., 2016) investigaram o desempenho do *Hadoop* para cargas aplicadas por um conjunto de *benchmarks* em ambientes físicos e em ambientes de nuvem privada(Openstack). A carga de trabalho considerada é o tamanho do *dataset* e as variáveis de resposta contempladas foram o tempo de execução do *Hadoop job* assim como o consumo de processamento de máquinas virtuais e máquinas físicas configuradas para analisar os conjuntos de dados. Nesse trabalho, a aplicação *Terasort* apresentou o maior tempo de execução comparado ao *Teravalidate* e o *Teragen*. O estudo também demonstrou que a utilização de processamento de máquinas virtuais em ambientes de nuvens comparados às arquiteturas nativas foi menor, cerca de 10% de vantagem, visto que máquinas virtuais apresentam uma utilização de recursos adequados à demanda gerada por cargas de trbalho e maior escalabilidade considerando diferentes tamanhos de *datasets*.

Lu et al. (LU et al., 2017) apresentaram uma avaliação de desempenho quando análises de *Big Data* são executadas por *Hadoop clusters* suportados por ambientes de nuvem. A geração de cargas de trabalho foram caracterizadas pelo TestDFSIO, que executa operações de leitura e escrita em disco(*I/O intensive jobs*), *Pi* e *Terasort*, que possuem mais ênfase em utilização de *CPU* da infraestrutura que suporta estas aplicações. Nos experimentos, a comparação de desempenho foi realizada entre um *cluster A* formado por 7 máquinas virtuais com 4 *cores* de processamento, 10 GB de memória e 100 GB de disco, e um *cluster B*, no qual um *host* físico hospeda 14 máquinas virtuais com 2 *cores* de processamento, 5 GB e 100 GB(metade da capacidade adotada no *cluster A*). Para os testes realizados com a escrita de dados, o *cluster A* apresentou tempos de execução cerca de 20% menores comparados ao *cluster B*, enquanto que para o teste de leitura de dados, a vantagem do *cluster A* em relação ao *cluster B*, foi de aproximadamente 48%. Para o teste que aplica cargas de trabalho no processador, o impacto da mudança do número de máquinas virtuais, assim como suas ofertas de serviços, foi de 2% na utilização de processadores no ambiente *Big Data*. Os experimentos demonstraram que operações de leitura e escrita em disco em altas ofertas de serviço em nuvens privadas proporcionam melhores resultados de desempenho comparados com ambientes de menores ofertas de serviço. Outra conclusão deste trabalho consiste no fato de que quando o recurso de processamento é avaliado em operações comuns em análise de *Big Data*, o tamanho do

cluster não apresenta impacto significativo no desempenho destas transações.

Chen e Rodero (CHEN; RODERO, 2017) caracterizaram o desempenho assim como aspectos de consumo de energia considerando dois ambientes de análise *Big Data*: O *Hadoop* em sua forma mais tradicional e o *Spark*, uma das ramificações do *Hadoop*. Cargas de trabalho geradas por operações como *Grep*, *Wordcount*, e *K-means*, foram executadas com o objetivo de analisar *datasets* tais como textos do Wikipédia, dados de redes sociais, dentre outros. A análise do consumo de energia do *Spark* foi menor para todas as operações de *Grep, K-means, e Wordcount* assim como a utilização de processador também se apresentou menor comparado à arquitetura padrão do *Hadoop*. Um algoritmo foi proposto para direcionar as cargas de trabalho considerando 10 *clusters*, cada um composto por 8 *data nodes*. Os resultados da simulação demonstraram que o fator alocação de banda de rede da nuvem apresentou impacto significativo em ambos os ambientes de análise *Big Data*.

3.2 Avaliação de Desempenho e Custo

Essa seção apresenta trabalhos que avaliam o desempenho e o custo de ambientes *Big Data* em infraestruturas de nuvem.

Eckroth (ECKROTH, 2016) realizou 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*. O custo associado à nuvem pública se apresentou menor em relação ao custo relacionado à utilização de nuvem privada.

Ruiz et al. (RUIZ et al., 2016) apresentaram uma avaliação de desempenho e custo de *MapReduce* em plataformas de nuvem, avaliando *trade-offs* como número de data nodes versus tempo de execução e custo de recursos. Adicionalmente, redes de Petri foram adotadas para modelar o *MapReduce* em ambiente de nuvem através de análise do espaço

de estados e avaliação de desempenho. A rede de Petri se baseou em quatro principais atividades de um *Hadoop job*, cada atividade com um tempo associado em milissegundos. A primeira tarefa é o tempo necessário para preparar o sistema, como colocar arquivos no *HDFS*, instanciar máquinas virtuais que irão fazer parte do *Hadoop cluster*, como um *master nodes, secondary name nodes, e data nodes*. A segunda tarefa do *job* é o *Map*, no qual os pares chave-valor(*key/value)* são formados. A terceira tarefa consiste no *Reduce*, em que os pares chave-valor são processados com o intuito de obter os resultados desejados pela análise dos *datasets*. A última tarefa é chamada de *Cleanup*. Nesta atividade, o *HDFS* armazena os resultados da análise e elimina arquivos intermediários. O modelo de desempenho avaliou os custos associados à implantação de *data nodes* no *Hadoop cluster* para a análise de sentimento de 100 milhões de *posts* na rede social *Twitter*. Este artigo mostrou que o custo total de 7 horas de tempo de execução considerando 2 *data nodes* é em média $43,92, enquanto que o aumento do número de *data nodes*, resultou em um menor tempo de execução assim como um menor custo comparado ao cenário com uma menor quantidade de *data nodes*, demonstrando a importância de fatores como número de *data nodes* para critério de desempenho de *Hadoop clusters* suportados por infraestruturas de nuvens privadas.

Lovas et al. (LOVAS et al., 2018) apresentaram uma interface para implementação do *Occopus cloud orchestrator*, no qual medições de desempenho e custo foram avaliadas. A primeira avaliação foi baseado no tempo de implantação(*deployment time* demandado por um *Hadoop cluster* considerando de 4 as 12 *data nodes* e três diferentes plataformas de nuvem: *MTA cloud*, baseado em Openstack; *LPDS Cloud Based*, baseado em OpenNebula; e Cloudsigma. A plataforma *MTA Cloud* apresentou os melhores tempos de *deployment* considerando de 4 a 12 *data nodes*. A avaliação de custo do ambiente *Hadoop* demonstrou que o custo pode ser reduzido com a implantação da interface proposta pelos autores. O *Occopus* apresentou reduções de custo consideráveis comparados à serviços de *big data* em nuvem como o *HDInsight*. Para a maior oferta de serviço adotada(maior capacidade), a redução de custo chega a ser de 3000 Euros aproximadamente anuais.

3.3 Comparação dos Trabalhos Relacionados

A Tabela 1 denota as contribuições dos trabalhos relacionados a esta dissertação. Estes trabalhos são comparados em relação a apresentação de uma metodologia para avaliação de desempenho ou avaliaçao de custo (Met); da técnica de medição para avaliação de desempenho ou custo (Med); de modelos de desempenho (MD), de modelos de custo (MC) e do planejamento de experimentos (PE). Pode-se notar que todos os trabalhos adotaram a técnica de medição para a avaliação de desempenho.

Tabela 1 – Tabela comparativa de trabalhos relacionados

Trabalho	**Met**	**Med**	**MD**	**MC**	**PE**
(RUIZ et al., 2015)	x	x	x	x	
(ZHANG; SAKR, 2014)		x		x	
(MASSOBRIO et al., 2018)		x			x
(KIM; HUH, 2017)	x	x	x		
(ADHIKARI et al., 2017)		x			x
(VELLAIPANDIYAN; SRIKRISHNAN, 2014)		x			x
(FEMMINELLA et al., 2016)		x			x
(LU et al., 2017)		x			x
(CHEN; RODERO, 2017)		x			
(ECKROTH, 2016)		x		x	
(RUIZ et al., 2016)		x	x	x	
(LOVAS et al., 2018)		x		x	
Esta dissertação	x	x	x	x	x

Apenas dois trabalhos relacionados apresentaram uma metodologia para a avaliação de desempenho e/ou custo de aplicações *Big Data* em nuvens privadas. Três trabalhos apresentaram modelos de desempenho para avaliar ambientes *Big Data* em infraestruturas de nuvens privadas e quatro apresentaram modelos de custo para estes ambientes. Este trabalho de dissertação visa a avaliação conjunta de desempenho e custo de ambientes *Big data* suportados por infraestruturas de nuvens privada e abordou as diferentes técnicas e a proposição de metodologia, com o objetivo de prover uma avaliação mais completa, contemplando diferentes aspectos que os trabalhos relacionados não contemplaram.

3.4 Considerações Finais

Este capítulo trouxe os principais trabalhos relacionados a este documento, em que uma pesquisa sistemática em bases científicas relevantes foi realizada com o objetivo de selecionar os trabalhos com maior afinidade. Muitos trabalhos apresentaram apenas medições de desempenho aliados ao planejamento de experimentos para avaliar aplicações *Big Data* em ambientes de nuvem. Por outro lado, nenhum trabalho apresentou experimentos com simulações e poucos apresentaram modelos de desempenho e custo. A metodologia foi outro ponto não observado na maioria dos artigos. O seguinte capítulo apresenta a metodologia para a avaliação de desempenho e de custo de ambientes *Big Data* implantados em infraestruturas de nuvens privada.

4 Metodologia para Avaliação de Desempenho e Custo de Ambientes *Big Data* em Infraestruturas de Nuvens Privadas

Este capítulo apresenta a metodologia proposta para avaliar o desempenho e o custo de ambientes *Big Data* implantados em infraestruturas de nuvens privadas.

Diversos ambientes de *Big data* como o *Apache Spark*, *Cloudera* e *Hortonworks*, podem adotar a metodologia para avaliar o desempenho e custo ao realizar a análise de grandes conjuntos de dados. O requisito para estes outros ambientes, é que estejam implantados em ambientes de nuvem privada, no qual o usuário possui um maior nível de gerenciamento quanto aos recursos utilizados e custos associados.

A visão geral da metodologia aborda as principais atividades e em seguida, cada uma delas são descritas de forma mais aprofundada. A metodologia proposta contempla a geração de modelos de desempenho baseados em redes de Petri estocástica, a geração de modelos de custo, a identificação de fatores impactantes para o desempenho e o custo e a análise de novos cenários considerando as métricas coletadas em ambientes reais.

4.1 Visão Geral da Metodologia Proposta

A metodologia proposta visa avaliar o desempenho e o custo de ambientes *Big Data* em nuvens privadas através de um modelo de desempenho baseado em redes de Petri estocásticas e modelos de custo baseados em equações matemáticas. A metodologia ainda contempla a análise de sensibilidade de fatores impactantes no gerenciamento da infraestrutura da nuvem privada, como utilização dos recursos de processamento e memória de máquinas virtuais e custos da aplicação *Big Data* implantada em nuvem privada.

A Figura 9 denota uma visão de alto nível da metodologia proposta. O entendimento do ambiente compreende a definição dos requisitos de *hardware* e *software* necessários para configuração do ambiente *Big data* na nuvem privada. Equipamentos da infraestrutura de nuvem privada e as aplicações devem ser definidas de acordo com o objetivo da avaliação de desempenho e de custo. O planejamento de experimentos consiste na definição de fatores e níveis para avaliar o impacto e analisar estatísticamente as métricas adotadas. A medição consiste na coleta das métricas de desempenho do consumo de energia elétrica considerando

diferentes níveis dos fatores adotados para os experimentos. A modelagem de desempenho consiste em representar aplicações *Big data* configuradas em nuvens privadas através de redes de Petri estocásticas. A modelagem de custo avalia o custo da infraestrutura de nuvem necessária para configurar o ambiente *Big Data*, o custo do consumo energético desta infraestrutura e da aplicação *Big data* e o custo de *software* adquirido para executar a análise dos conjuntos de dados.

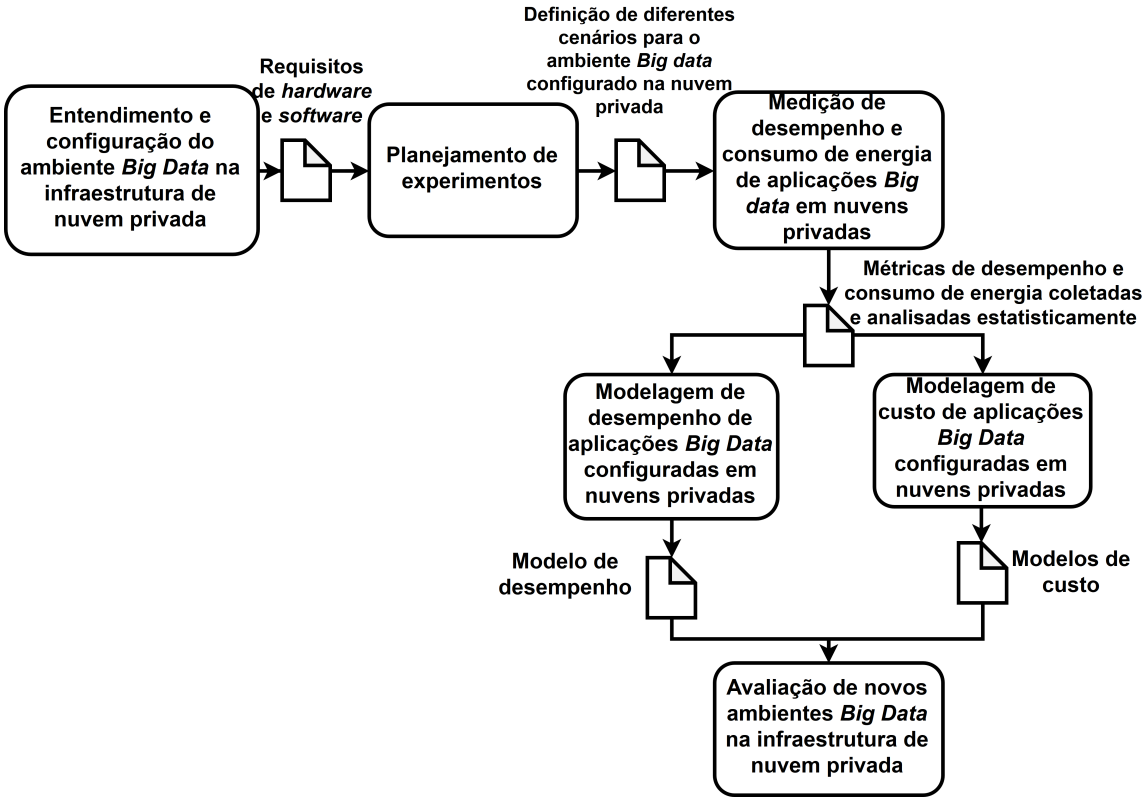

Figura 9 – Visão Geral da Metodologia para Avaliação de Desempenho e Custo de Ambientes Big Data em Nuvens Privadas (FONTE: Próprio Autor)

Na saída de cada atividade da metodologia em alto nível, há um artefato que ilustra o objetivo e resultados da atividade. Por fim, novos cenários podem ser analisados com o auxílio do modelo de desempenho e custo validados.

A Figura 10 apresenta os elementos adotados para representar a metodologia proposta. O elemento atividade representa ações que devem ser executadas na sequência lógica definida pela metodologia. O artefato indica o *output* ou saída de cada atividade. O elemento de decisão denota o controle da sequência de atividades, enquanto as ligações entre as atividades com sequências lógicas de fluxo são representadas pelos elementos de conexão entre elas (CHINOSI; TROMBETTA, 2012).

Analistas de desempenho, analistas de T.I. e gerentes de empresas de *data analytics*

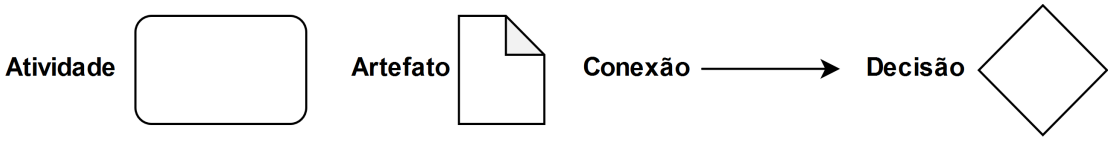

Figura 10 – Elementos da Metodologia Proposta (CHINOSI; TROMBETTA, 2012)

podem adotar a metodologia proposta para ser aplicada em diferentes ambientes *Big Data* configurados em infraestruturas de nuvens privadas. As atividades relacionadas ao planejamento de experimentos e modelagem de desempenho podem ser realizadas por analistas de desempenho devido à necessidade do conhecimento das técnicas de planejamento de experimentos e do formalismo de redes de Petri estocásticas.

O entendimento do ambiente, a medição de desempenho e a modelagem de custo podem ser realizadas por analistas de tecnologia da informação. Gerentes de empresas de *data analytics* e *data analysis* podem se basear na metodologia proposta para avaliar o desempenho e custo de ambientes *Big Data* em infraestruturas de nuvens privadas, considerando diferentes recursos de processamento e memória das máquinas virtuais, quantidades de *datanodes* e tamanho dos *data sets*.

4.2 Atividades da Metodologia para Avaliação de Desempenho e Custo de Ambientes *Big Data* em Nuvens Privadas

Neste contexto, são apresentadas a seguir, as atividades contempladas na metodologia proposta, são elas: Entendimento do ambiente de análise *Big Data* baseado em nuvem privada; Seleção de métricas de desempenho e consumo de energia elétrica em ambientes *Big Data*; Planejamento de Experimentos para o desempenho da aplicação *Big Data* em infraestruturas de nuvens privadas; Geração da carga de trabalho *Big Data*; Execução de aplicações *Big Data* em infraestruturas de nuvens privadas; Análise estatística das métricas de desempenho; Geração do modelo de desempenho e consumo de energia elétrica; Avaliação das propriedades do modelo de desempenho; Refinamento do modelo de desempenho; Mapeamento de métricas de desempenho; Validação do modelo de desempenho e Análise de novos cenários.

Modelar o sistema considerando aspectos de desempenho e custo é essencial para avaliação de nuvens privadas executando aplicações *Big Data*. Modelos *SPNs* e equações de custo, contribuem para estimar cenários mais robustos, que geralmente são mais complexos

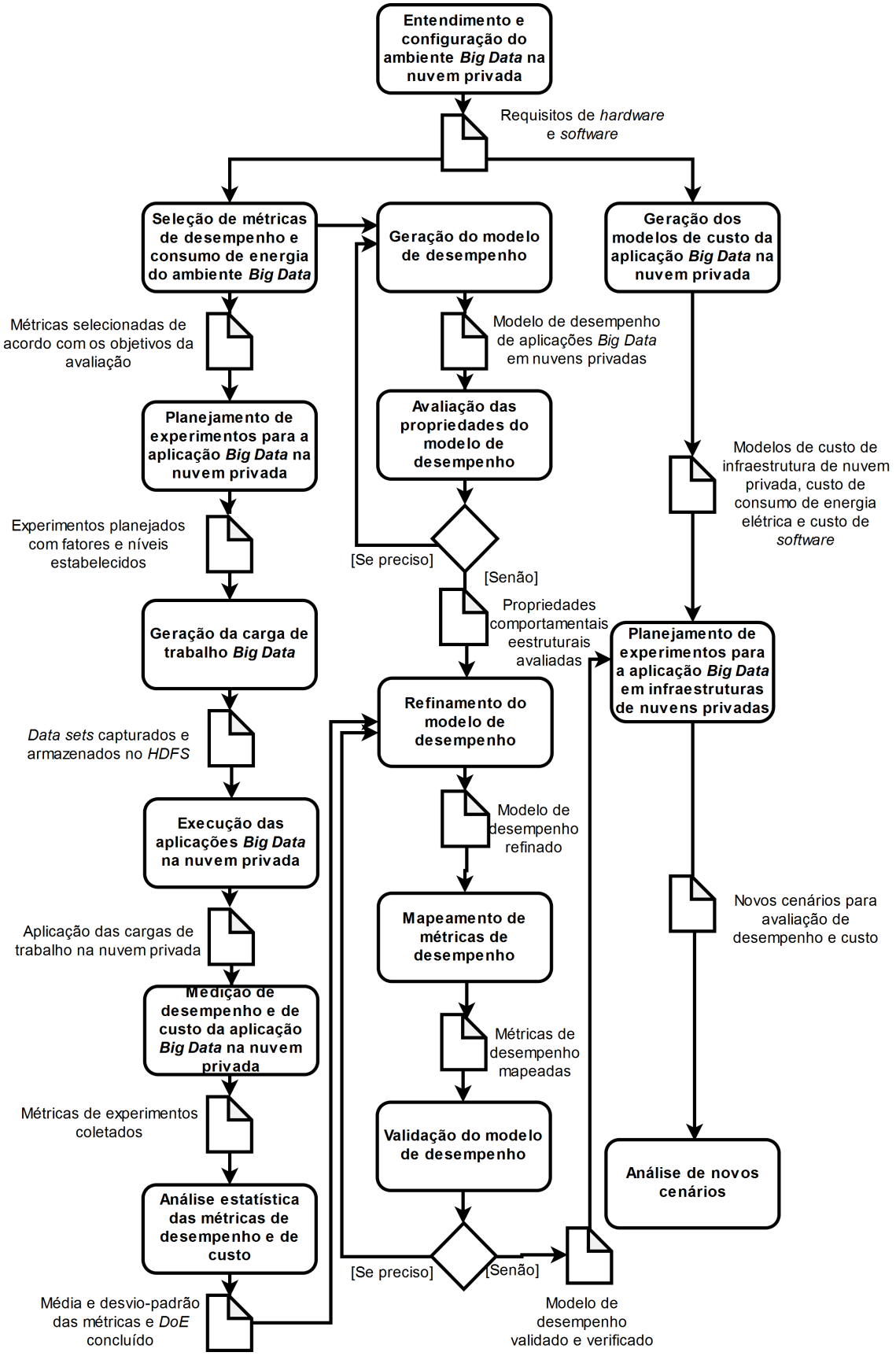

Figura 11 – Metodologia para Avaliação de Desempenho e Custo de Ambientes *Big Data* em Nuvens Privadas (FONTE: Próprio Autor)

de implantar em ambientes reais. Um exemplo é considerar a instanciação de 100 máquinas virtuais para analisar Petabytes de dados. Isto pode demandar um longo período de tempo assim como uma grande necessidade de recursos. A Figura 11 apresenta a Metodologia para Avaliação de Desempenho e Custo de Ambientes Big Data em Nuvens Privadas.

4.2.1 Entendimento e configuração do ambiente de análise *Big Data* em nuvem privada

Administradores de sistemas devem identificar os componentes de *software*, *hardware* e fatores do sistema que possuem capacidade de afetar o desempenho e o custo de aplicações *Big Data*. Todas as configurações da nuvem privada e de suas máquinas virtuais são estabelecidas com base no planejamento de experimentos, consdirandos ofertas de serviços de máquinas virtuais instanciadas e o número de *data nodes* que compõem o *Hadoop cluster* (ALAPATI, 2016).

É importante avaliar as plataformas de nuvem privada que podem ser adotadas para provisionar os recursos computacionais para as transações *Big data*. Dentre as principais plataformas de nuvem privada, temos o Apache Cloudstack (CLOUDSTACK, 2018) e OpenStack (OPENSTACK, 2017) e Eucalyptus (EUCALYPTUS, 2017). O ambiente de nuvem deve ser selecionado considerando as plataformas de nuvens privadas, pois permite controle sobre a infraestrutura de nuvem privada como gerenciamento de desempenho de *hosts*, ofertas de serviços e capacidade total disponível para a instanciação das máquinas virtuais. As plataformas de nuvem privada oferecem o *pool* de recursos computacionais que serão consumidos durante a execução das cargas de trabalho geradas por aplicações *Big Data*. Essa atividade também considera a identificação das métricas desempenho e consumo de energia elétrica adotadas na avaliação da aplicação *Big Data* na nuvem privada. A capacidade da nuvem privada define a quantidade de máquinas virtuais que podem ser instanciadas.

Dentro deste contexto, mecanismos de análise de dados surgiram da crescente produção de dados gerados pela sociedade e suas atividade, sendo obtidos por diferentes fontes tais como redes sociais, dados médicos, e dados climatológicos por sensores (HASHEM et al., 2015; ASSUNÇÃO et al., 2015). As máquinas virtuais da nuvem privada são instanciadas para analisar os *data sets* baseado em requisições geradas por

algoritmos enviados por clientes que possuem grandes quantidades de dados para serem analisados. Diferentes quantidades de máquinas virtuais podem ser instanciadas de modo a oferecer mais capacidade e replicabilidade para a execução de aplicações *Big Data*.

4.2.2 Seleção de métricas de desempenho e consumo de energia em ambientes *Big Data*

Após o entendimento do ambiente *Big Data*, os analistas de desempenho devem escolher as métricas para a avaliação de desempenho e consumo de energia dos serviços *Big Data* executados na nuvem privada considerando diferentes *data sets* e ofertas de serviço. O monitoramento e análise das métricas de desempenho e consumo de energia permitem que os administradores de sistemas verifiquem os níveis de qualidade oferecidos aos clientes e um maior controle da infraestrutura, sendo escalável para suportar cargas de trabalho com variadas intensidades. Na literatura (MARINESCU, 2017), métricas que representam qualidade de serviços de computação em nuvem são ilustradas por tempos de resposta, tempos de execução (YADAV et al., 2018), e vazão de requisições de usuários, enquanto métricas de utilização de recursos representam aspectos de gerenciamento da infraestrutura de nuvem.

4.2.3 Planejamento de Experimentos para a aplicação *Big Data* em infraestruturas de nuvens privadas

Definidas as métricas que serão consideradas para a avaliação de desempenho e custo de aplicações *Big Data* em infraestruturas de nuvens privadas, deve-se planejar de forma sistemática os experimentos a serem executados, considerando a capacidade oferecida às máquinas virtuais implantadas na infraestrutura de nuvem. Esta atividade visa também identificar e determinar quais fatores têm mais impacto nas métricas de desempenho, consumo de energia e de custo quando as cargas de trabalho geradas pela aplicação *Big Data* são enviadas.

Fatores são entidades do problema que têm a capacidade de perturbar as métricas de desempenho e de custo, que variam em um intervalo controlado para avaliar seu impacto em variáveis de reposta (MONTGOMERY, 2017). Uma vez definidos os fatores e seus

níveis, a análise do *DoE* fornece informações importantes, como a hierarquização dos fatores adotados considerando seus impactos em variáveis de resposta(tempo de execução de experimentos, utilização de recursos e custo da aplicação *Big Data*).

Uma vez concluído o planejamento de experimentos, deve-se implantar a arquitetura formada por máquinas virtuais instanciadas na nuvem privada, de modo a fornecer os recursos computacionais necessários para suportar as cargas de trabalho *Big Data*.

4.2.4 Geração da carga de trabalho *Big Data*

O objetivo desta atividade é capturar, manipular e tratar os conjuntos de dados obtidos de diversas fontes. Com os experimentos planejados e a arquitetura implantada, analistas de desempenho podem gerar a carga de trabalho que será submetida e processada. Conjuntos de dados podem ser obtidos de base de dados ou capturados de diferentes fontes, como redes sociais, dados de saúde(*healthcare data*) e dados originados de sensores (HASHEM et al., 2015; ASSUNÇÃO et al., 2015; OUSSOUS et al., 2018).

Fontes de dados podem ser encontradas na comunidade, como é o caso de conjuntos de dados fornecidos para a comunidade publicamente. Por outro lado, o gerente de *data analysis* pode gerar os dados médicos, de redes sociais e a partir de métricas de sensores que podem ser captados e analisados de forma manual por analistas de *Big Data*. Estes são analisados pelo *cluster* de máquinas virtuais implantadas no ambiente de nuvem privada e são geradas informações acerca do conjunto de dados analisado. Está análise pode se referir a encontrar padrões de sequência no *DNA* humano (WANG et al., 2018) assim como analisar o sentimento de comentários em redes sociais considerando palavras ou frases mais citadas (HASHEM et al., 2015; ASSUNÇÃO et al., 2015).

A equipe de *data analytics* precisa entender como os conjuntos de dados serão coletados, capturados, manipulados, processados e analisados. Fluxogramas podem ser adotados para avaliar como ocorre o fluxo de geração da carga de trabalho, desde sua captura até a o seu envio para ser analisado na nuvem privada.

4.2.5 Execução de aplicações *Big Data* em infraestruturas de nuvens privadas

Com os *data sets* obtidos, aplicações *Big Data* são executadas na infraestrutura de nuvem privada através da instanciação de máquinas virtuais para analisar o *data set* e assim, gerar a carga de trabalho para cada cenários definidos no planejamento de experimentos, com seus fatores e níveis estabelecidos. Algoritmos são responsáveis por gerar as solicitações e é oferecido por empresas de *data analytics*, como é o caso de serviços tais como encontrar palavras mais citadas em redes sociais e identificar padrões de usuários em sistemas de saúde. Funções como agregar diversos *data sets*, ordenar de acordo com critérios como popularidade e geolocalização, são exemplos de requisições em ambientes *Big Data* (ASSUNÇÃO et al., 2015; HASHEM et al., 2015). O algoritmo deve ser desenvolvido com base nas características do ambiente escolhido para ser avaliado, tais como dados de redes sociais, de sensores, e de saúde.

4.2.6 Medição de desempenho da aplicação *Big Data* na nuvem privada

Uma vez obtidos os conjuntos de dados a serem adotados como carga de trabalho, a equipe de análise de *Big Data* deve armazenar os conjuntos de dados no sistema de armazenamento da nuvem privada. Mecanismos de tolerância a falhas e divisão de carga de trabalho são consideradas no armazenamento do conjunto de dados (ALAPATI, 2016).

Nesta atividade, o analista de desempenho deve definir o intervalo de tempo de medição e o intervalo de amostragem para coletar as métricas (LILJA, 2005). Além disso, o ambiente de nuvem privada e as máquinas virtuais devem ser reinicializadas após cada experimento.

As métricas devem ser coletadas de acordo com o planejamento de experimentos e suas configurações. Depois de executar as ferramentas de monitoramento, como o SYSSTAT (JUVE et al., 2015) e PERFMON (WINDOWS, 2018), analistas de desempenho e do ambiente *Big Data* podem enviar a solicitação ao *cluster* do *Hadoop*, executando o algoritmo desenvolvido, que coordena os *data nodes* responsáveis por provisionar os recursos para a análise dos dados.

A lista a seguir ilustra o procedimento a ser realizado para as coletar métricas de desempenho e custo de máquinas virtuais executando transações *Big Data*, sendo suportadas por infraestruturas de nuvens privadas.

1. Reiniciar todas as máquinas virtuais na nuvem privada e iniciar novos experimentos;
2. Em seguida, um *warm up* de unidades de tempo deve ser adotado antes de aplicar a carga de trabalho, como por exemplo, 10 segundos;
3. Depois do *warm up*, deve-se executar ferramentas de monitoramento de desempenho como o Sysstat (JUVE et al., 2015) e o Perfmon (WINDOWS, 2018) até o fim o término do intervalo de medição que contempla um tempo de *cooldown*. Após o intervalo de tempo de *warm up*, a carga de trabalho, caracterizado por algoritmos que estabelecem a dinâmica das transações *Big Data* são submetidas;
4. Ao encerrar a aplicação *Big Data* geradora da carga de trabalho sobre a infraestrutura de nuvem privada, pode-se então encerrar os *scripts* de monitoramento, após um intervalo de tempo de *cooldown*. Estes *logs* são gerados pelos *scripts* de monitoramento no formato *.txt* e podem ser exportados para planilhas de cálculos como Excel (WINSTON, 2016) e Minitab (MINITAB, 2018).
5. Os *logs* gerados por ferramentas de monitoramento com as métricas de desempenho e consumo de energia para cada cenário contemplado no planejamento de experimentos são armazenados em um computador que não foi utilizado na configuração da nuvem privada. Esses *logs* são gerados em formatos *.txt*, *.csv*, dentre outros.
6. Iniciar novos experimentos repetindo este procedimento a partir da Atividade 1.

Os experimentos são replicados de acordo com o número definido pelo analista de desempenho(n) de forma a proporcionar uma maior confiabilidade dos resultados acerca do desempenho e consumo de energia das aplicações *Big Data* executadas em nuvens privadas. O ambiente de medição adotado para os experimentos devem seguir as configurações adotadas no planejamento de experimentos. Capacidades de máquinas virtuais, sistemas operacionais, e capacidade da nuvem privada devem ser consideradas assim como suas variações de forma sistemática e alinhada.

A infraestrutura de nuvem privada oferece o *pool* de recursos para que a aplicação *Big Data* seja executada com base nas requisições de clientes, que possuem *data sets* a serem analisados e requisitam máquinas virtuais implantadas na nuvem privada. Atividades como estabelecer o intervalo de medição, tempo de amostragem das métricas de desempenho e o isolamento do ambiente devem ser contemplados nas atividades de medição de forma a obter métricas sem a interferência da execução de uma carga de trabalho externa.

4.2.7 Análise estatística das métricas de desempenho e consumo de energia

Para cada métrica de desempenho e consumo de energia adotada, deve realizar a análise estatística dos dados providos pelos experimentos. A primeira tarefa desta atividade é remover *outliers* das séries de dados relacionados às métricas. *Outliers* podem levar a distorções nos resultados das métricas (GUPTA; GUTTMAN, 2014). Após esta remoção, o cálculo da média e do desvio padrão das métricas de desempenho e consumo de energia são realizadas para os cenários adotados, em que os resultados das métricas de desempenho e consumo de energia são avaliados através da execução da análise de projeto fatorial (MONTGOMERY, 2017) com o objetivo de estabelecer os principais fatores que afetam o desempenho, consumo de energia e custo de transações *Big Data* em nuvens privadas além de prover as médias e desvios-padrões para cada cenário.

4.2.8 Geração dos modelos de custo da aplicação *Big Data* na nuvem privada

Esta atividade tem como objetivo avaliar o custo da implantação de ambientes *Big Data* em infraestuturas de nuvens privadas. Componentes que impactam no custo do ambiente *Big dat* devem ser considerado para definir os diferentes componentes do custo total para realizar a análise de grandes conjuntos de dados em nuvem privada. Custos de componentes da nuvem privada como máquinas físicas, *switches* e roteadores denotam o custo de infraestrutura da nuvem privada. O custo do consumo de energia é outro componente de custo que tem como fatores o tempo de execução da aplicação *Big Data* e o consumo de energia que o ambiente demanda para a análise do *data set*. Custo de aquisição de *software* como plataformas de nuvem privada e aplicações *Big Data* devem ser contemplados nesta atividade. O custo total do ambiente *Big Data* é proposto neste trabalho como a soma dos custos de infraestrutura de nuvem privada, custo associado ao consumo de energia e custo de aquisição de *software* para implantação do ambiente.

4.2.9 Geração do modelo de desempenho

Essa atividade tem o objetivo de gerar o modelo de desempenho baseado em redes de Petri estocásticas para representar o processamento de conjuntos dados na nuvem privada. Esse modelo de desempenho contempla a representação do cliente enviando a

requisição de análise *Big data* além de aspectos como diferentes capacidade da nuvem privada e tamanhos de conjuntos de dados. O modelo de desempenho prosposto considera a representação de diferentes níveis de carga de trabalho, diferentes tamanhos de datasets e os recursos de memória e processamento usados dos *data nodes* para o processamento dos *data sets*. O modelo de desempenho é adotado para avaliar métricas de utilização de recursos em máquinas virtuais que tem a função de analisar grandes quantidades de dados.

4.2.10 Avaliação das propriedades do modelo de desempenho

Uma vez que o modelo de desempenho para avaliar o desempenho do ambiente *Big Data* é gerado, é necessário avaliar as suas propriedadades. Esta análise qualitativa do modelo proporciona a avaliação das propriedades comportamentais e estruturais do modelo de desempenho tais como alcançabilidade, vivacidade, limitação e presença de *deadlocks* (MURATA, 1989; MACIEL et al., 2017) através de ferramentas como o HiPS tool (HIPS, 2018) e INA (INA, 2018). No processo de análise das propriedades, pode-se observar a necessidade de ajustes no modelo de desempenho e caso alguma propriedade avaliada não seja satisfeita, o modelo deve ser gerado novamente.

4.2.11 Refinamento do modelo de desempenho

Esta atividade fornece o modelo de desempenho renado com base no inverso do coeciente de variação $(1/Cv)$, calculado de acordo com as médias e desvios-padrões dos tempos de mapeamento e de redução do processamento do conjunto de dados na nuvem privada (YEE; VENTURA, 2000).

De acordo com Feitelson (FEITELSON, 2015), os tempos de mapeamento e de redução podem não seguir o comportamento de uma distribuição de probabilidade exponencial. Neste caso, a aproximação de fases (YEE; VENTURA, 2000) indicará a distribuição de probabilidade expolinomial que melhor representa a distribuição de probabilidade empírica dos tempos de mapeamento e de redução do *data set* analisado.

O renamento do modelo de desempenho é realizado neste trabalho pela técnica de aproximação de fases, que calcula o primeiro (μ_D) e segundo momentos (σ_D) da distribuição de probabilidade empírica dos tempos de mapeamento e de redução. Esse cálculo provê a

seleção da distribuição de probabilidade expolinomial e os parâmetros numéricos desta distribuição de probabilidade.

4.2.12 Mapeamento de métricas de desempenho

As métricas de desempenho precisam ser implementadas no modelo de desempenho e devem representar as métricas definidas na avaliação de desempenho. O objetivo dessa atividade é representar o conjunto de critérios de desempenho de aplicações *Big Data* em nuvens privadas em métricas através de referências aos elementos do modelo de desempenho refinado concebido.

4.2.13 Validação do modelo de desempenho

Esta atividade compara as métricas coletadas das medições do ambiente *Big Data* na nuvem privada com métricas obtidas no modelo de desempenho refinado. Os valores de cada cenários dos experimentos reais devem ser comparados nos mesmos cenários considerando os resultados provisionados pelo modelo de desempenho refinado. Estes resultados devem ser equivalentes com uma porcentagem aceitável de erro(95% de confiança (GUPTA; GUTTMAN, 2014) ou de entre 10% a 20% de erro relativo (MENASCE et al., 2004)). Recomenda-se adotar o teste T emparelhado ou a validação percentual para este fim (GUPTA; GUTTMAN, 2014). A validação do modelo de desempenho é necessária para avaliar cenários não abrangidos nas medições, como considerar altas intensidades de carga de trabalho assim como considerando um número maior de *data nodes* para a análise de dados. Caso o modelo de desempenho refinado não seja validado, será necessário realizar ajustes.

4.2.14 Planejamento de experimentos para a aplicação *Big Data* na nuvem privada

Esta atividade tem como objetivo realizar o planejamento de experimentos, que define novos fatores e níveis a serem contemplados na análise de novos cenários. Diferentes capacidades do ambiente *Big data* e tamanhos de *data sets* são avaliados com base nos

modelos de desempenho refinado e de custo gerados anteriormente. O novo planejamento de experimentos define novas replicações e execuções para avaliar estes cenários.

4.2.15 Análise de novos cenários

Esta atividadade tem o objetivo de proprocionar a análise de ambientes não contemplados nos experimentos adotando o modelo de desempenho e os modelos de custos proposto neste trabalho. Nessa atividade, o modelo de desempenho validado e refinado é adotado para realizar a análise das métricas de desempenho e de consumo de energia considerando fatores como a análise de maiores tamanhos de *data sets* e maiores capacidades do ambiente *Big data*. O resultado da atividade são a avaliação de desempenho e de custo de novos ambientes *Big data* através dos modelos e sem a necessidade de requisitos de *hadtrware* e *software* que estes novos cenários poderiam incorrer.

4.3 Considerações Finais

Este capítulo apresentou a sequência de atividades contempladas para a avaliação de desempenho e custo de infraestruturas de nuvens privadas para aplicações *Big Data*. O seguinte capítulo apresenta os modelos adotados nesta dissertação para avaliação de desempenho e custo de ambientes *Big Data* suportados por infraestruturas de nuvens privadas.

5 Modelo de Desempenho e Modelos de Custo

Este capítulo apresenta os modelos de desempenho e os modelos de custo para ambientes *Big data* configurados em infraestruturas de nuvens privadas. O modelo de desempenho é baseado no formalismo de redes de Petri estocásticas (MARSAN; BALBO, 1994; MARINESCU, 2017) e é adotado para avaliar a utilização de recursos dos *data nodes* do *cluster Hadoop* configurados nas máquinas virtuais da nuvem privada. Os modelos de custo de infraestrutura, de consumo de energia elétrica e custos de *software*, avaliam o custo de implantação de ambientes *Big Data* em plataformas de nuvem privada.

5.1 Modelo de desempenho de ambientes *Big data* em nuvens privadas

Aplicações realizadas em ambientes de nuvem podem ser modeladas por redes de Petri estocásticas, em que modelos *cliente/servidor* (JAIN, 1991; MENASCE et al., 2004; FEITELSON, 2015) 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 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. Estes dados são processados em duas fases principais: (1) o mapeamento dos dados; e (2) a redução dos dados. O mapeamento de dados consiste em distribuir o *data set* em pares chave-valor, no qual estes pares são distribuídos entre o *data nodes* do *Hadoop cluster*, neste caso configurados em VMs instanciadas na plataforma de nuvem privada. A redução dos dados consiste em processar os fragmentos do *data sets* formados por pares chave-valor para a obtenção do resultado esperado da análise de dados. A redução dos dados também é responsável por armazenar os resultados no sistema de armazenamento distribuído do *Hadoop*, o *HDFS* (ALAPATI, 2016). O objetivo do modelo é representar este processo formado pelo mapeamento e redução de *data sets* de forma temporizada, considerando a utilização de recursos tais como processador e memória.

A Figura 12 denota um exemplo para o entendimento de mecanismos *MapReduce* (ALAPATI, 2016). Neste caso é considerado um *data set* formado por cidades do nordeste e o objetivo é avaliar quantas palavras são referenciadas para cada cidade. A fase *Split* é uma fase intermediária que fragmenta o *data set* em conjuntos menores de dados para serem distribuídos pelo *Hadoop cluster*. A fase de mapeamento gera os pares chave-valor de interesse, armazenando quantas vezes a palavra é citada em cada fragmento do *data set*. O *Shuffling* é uma outra fase intermediária que quantifica o número de vezes que cada palavra foi citada em todo o *data set*. Por fim, a fase de redução dos dados obtém as informações de interesse e as guarda no *HDFS* (HASHEM et al., 2015; ALAPATI, 2016).

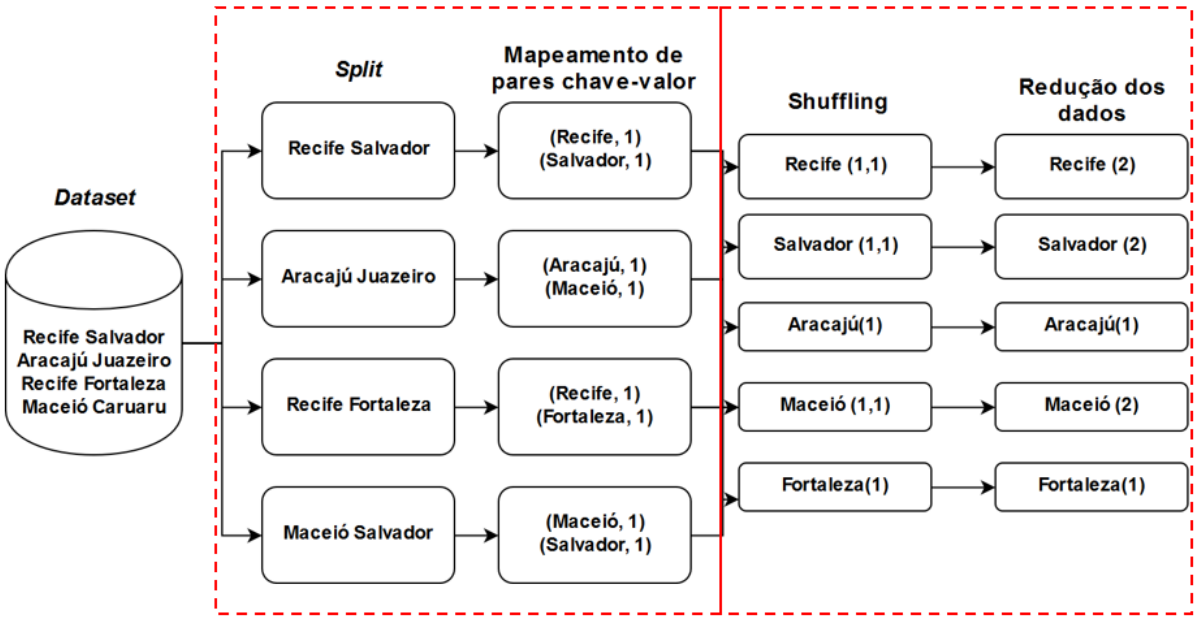

Figura 12 – Mecanismo de análise *Big Data* com *Hadoop clusters* FONTE: Próprio autor

O modelo de desempenho é baseado no formalismo de redes de Petri estocásticas, em que os tempos representados são os tempos para mapear e reduzir *data sets*, que podem ser gerados por diferentes fontes e dispositivos. O modelo de desempenho é composto por duas subredes principais: *Carga de trabalho* e *Cluster configurado na nuvem privada*. A Figura 13 apresenta o modelo de desempenho proposto para avaliação de desempenho da aplicação *Big data* em infraestruturas de nuvens privadas.

Na primeira subrede, o objetivo é gerar a carga de trabalho a ser submetida para o ambiente, em que n denota o tamanho do *data set* que o usuário envia para ser analisado, em que diferentes tamanhos de *data sets* podem ser considerados no modelo de desempenho. (T_{env}) representa o tempo para que a requisição de análise dos conjuntos de dadaos cheguem

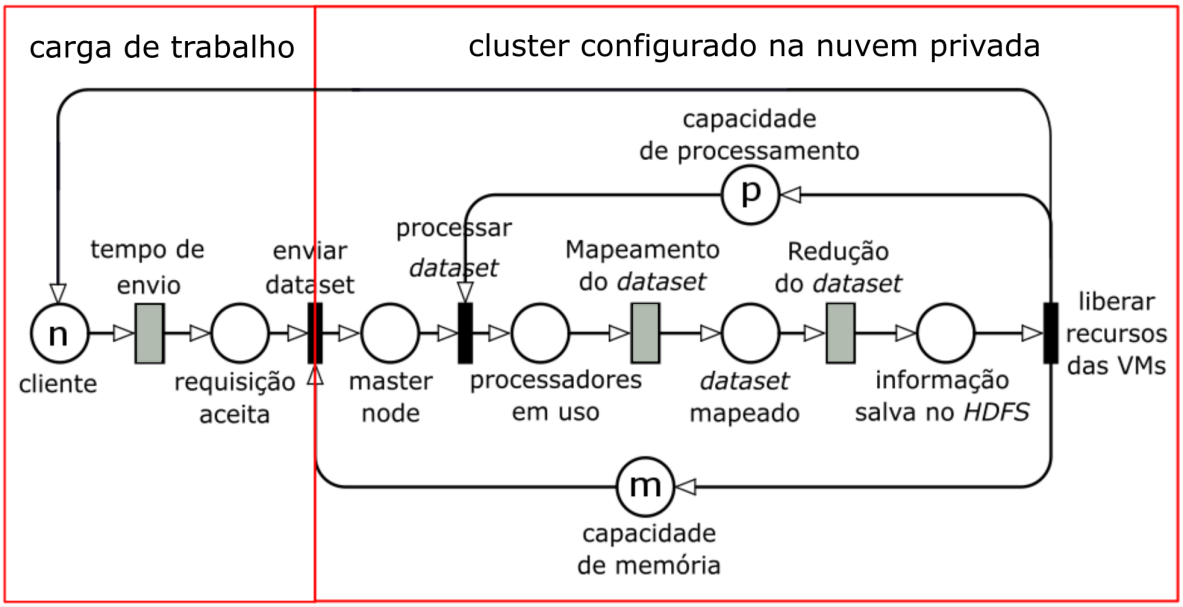

Figura 13 – Modelo de desempenho do hadoop cluster na nuvem privada (FONTE: Próprio autor)

à infraestrutura de nuvem privada, associado a transição *tempo de envio*. Após o disparo da transição temporizada *tempo de envio*, a requisição é aceita pelo *master node* e em seguida a transição *enviar dataset* é disparada. O *masternode* é responsável por coordenar o *cluster* quanto ao seu gerenciamento de recursos, além de armazenar os *data sets* a serem analisados.

Uma vez aceita a requisição, o *master node* a envia para os *data nodes* que armazenam em sua memória principal os dados a serem processados pelo *cluster* e armazena os procedimentos necessários para a análise do *data set*. Então, o conjunto de dados está pronto para ser processado, no qual há a utilização de recursos dos *data nodes*. O processamento de *data sets* consiste no mapeamento dos *data sets* assim como na sua redução e cada uma possui uma transição temporizada relacionada aos seus tempos de execução (transição *Mapeamento do dataset* e *Redução do data set*). Uma vez processado o *dataset*, a transição *liberar recursos das VMs* é disparada, liberando os recursos de processamento e memória dos *datanodes*.

Na subrede **carga de trabalho**, as marcações atribuídas ao lugar *cliente* definem a carga de trabalho que será enviada para o *Hadoop cluster*, em que o número de marcações é proporcional ao tamanho do *data set*. A transição **tempo de envio** está relacionada ao intervalo de tempo(T_{env}) em as requisições são enviadas para serem atendidas pelo *Hadoop cluster* implantados na nuvem privada. O lugar **requisição aceita** indica que o *data set*

será processado pelo *Hadoop cluster* após o disparo da transição **enviar dataset**.

Na subrede **carga de trabalho**, a marcação no lugar **masternode** representa a máquina virtual instanciada na nuvem privada onde o *master node* é configurado para coordenar as atividades de processamento do *data set*. Em seguida, a transição imediata **processar dataset** é disparada para denotar que o *data set* está sendo enviado para os *data nodes* do *Hadoop cluster*. A marcação no lugar **processadores em uso** indica que os processadores dos *data nodes* configurados na nuvem privada estão ocupados atendendo a requisição de análise do *data set*. O início do processamento do *data set* ocorre com o mapeamento do conjunto de dados, que demanda o tempo T_{map}, associado à transição **Mapeamento do dataset**. A marcação no lugar **dataset mapeado** denota que o conjunto de dados terminou o processo de mapeamento. A transição **Redução do dataset** representa o tempo demandado para a fase final da análise do conjunto de dados, a redução do *data sets*, que é associado ao tempo T_{red}. Uma vez reduzido o conjunto de dados, a marcação no lugar **informação salva no *HDFS*** indica que a informação desejado foi extraída do conjunto de dados. As marcações nos lugares **capacidade de processamento** e **capacidade de memória** denotam as capacidades do *Hadoop cluster*, em que as capacidades de cada *data node* são somadas para provisionar a capacidade total (ALAPATI, 2016).

Tabela 2 – Transições do modelo de desempenho proposto

Transição	**Tipo**	**Tempo**	**Peso**	**Prioridade**	**Concorrência**
TEMPO DE ENVIO	temp.	T_{env}	-	-	SS
ENVIAR DATASET	imed.	-	1	1	-
PROCESSAR DATASET	imed.	-	1	1	-
MAPEAMENTO DO DATASET	temp.	T_{map}	-	-	SS
REDUÇÃO DO DATASET	temp.	T_{red}	-	-	SS
LIBERAR RECURSOS DAS VMS	imed.	-	1	1	-

As transições podem ser imediatas (imed.) e temporizadas (temp.)(Tabela 2). O atributo *Tempo* apresenta o tempo associado à transição, no qual T_{map} representa o tempo para o mapeamento do conjunto de dados a serem processados pelo *Hadoop*, enquanto T_{red} denota o tempo associado para gerar as informações finais resultantes da análise do *data set*. O atributo peso refere-se à probabilidade da transição ser disparada em relação

às outras transições e o atributo prioridade define uma ordem na qual as transições devem ser disparadas durante a execução do modelo de desempenho. Por fim, SS(*single-server*) indica que a transição só pode ser disparada quando a marcação anterior é processada, com uma taxa constante. Por outro lado, IS(*infinite-server*) denota que a taxa da transição temporizada aumenta, sendo proporcional ao número de marcações em seu lugar de entrada (MARSAN; BALBO, 1994).

Os tempos T_{map} e T_{red} representam os tempos de mapeamento e redução em diferentes cenários com vários números de data sets, ofertas de serviço e cargas de trabalho. Esses tempos podem ser representados por distribuições de probabilidade diferentes da exponencial. A técnica de aproximação por fases pode ser adotada para identificar as distribuições de probabilidade que representam os tempos T_{map} e T_{red}. Uma vez definido os elementos do modelo de desempenho baseado em *SPNs*, este modelo deve ser avaliado quantos às suas propriedades estruturais e comportamentais (MARSAN; BALBO, 1994).

O cálculo do inverso do coeficiente de variação ($1/C_v$) da métrica medida permite a seleção de qual distribuiçãode probabilidade expolinomial que melhor representa seu comportamento (MACIEL et al., 2017). Neste trabalho, as distribuições de probabilidade adotadas para a aproximação por fases são as distribuições de probabilidade *Erlang*, *Hipoexponencial* e *Hiperexponencial*. Assim, quando $1/C_v$ é um número inteiro e diferente de um, a distribuição empírica deve ser caracterizada por uma distribuição de probabilidade *Erlang*. Quando $1/C_v$ é um número maior que um (mas não um inteiro), a distribuição empírica é representada por uma distribuição de probabilidade Hipoexponencial. Quando $1/C_v$ é um número menor que 1, a distribuição empírica deve ser representada por uma distribuição de probabilidade Hiperexponencial (YEE; VENTURA, 2000).

As Métricas mais adotadas para avaliar o desempenho de sistemas computacionais são tempos de resposta, tempos de execução e utilização de recursos tais como processador, memória, e disco de sistemas (JAIN, 1991; MENASCE et al., 2004; FEITELSON, 2015). As métricas adotadas no modelo de desempenho são apresentadas na Tabela 3: $Uprocessador$(utilização de processador) e $Umem$(utilização de memória) das máquinas virtuais instanciadas na nuvem privada para realizar a análise dos *data sets*. A utilização de disco no ambiente nos experimentos com o *Hadoop cluster* configurado na nuvem mostrou que a carga de trabalho requer pouca utilização deste recurso. Desta forma, as métricas utilização de processador e utilização de memória foram analisadas nesse trabalho.

Tabela 3 – Métricas de desempenho do modelo proposto

Métrica	**Expressão**
$Uprocessador$	$\frac{E\{\#processadores\ em\ uso\}+E\{\#dataset\ mapeado\}}{p} \times 100$
$Umem$	$(\frac{m-E\{\#capacidade\ de\ memoria\}}{m}) \times 100$

Uprocessador representa a fração de tempo que os processadores das máquinas virtuais onde os *data nodes* do *Hadoop cluster* estão configurados permanecem ocupados, realizando as atividades de mapeamento e redução do *data set* fornecido. *Umem* representa a fração da memória total das máquinas virtuais utilizada durante o processo de mapeamento e redução dos *data sets.*.

Considerando a métrica *Uprocessador*, *E{#processadores em uso}* denota o valor esperado do número de marcações no lugar *processadores em uso*, que indica que o início da fase de mapeamento dos *data sets* e *E{#dataset mapeado}* denota o término da fase de mapeamento do conjunto de dados. O parâmetro p totais representa a soma dos *cores* de processamento dos *data nodes* que compõem o *Hadoop cluster*. *Eexp* indica o número médio da expressão interna (*exp*), onde *exp* é #processadores em uso e #dataset mapeado.

Considerando a métrica *Umem*, *E{#capacidade de memoria}* indica fração de memória total utilizada para cada experimento. O parâmetro m representa a capacidade total de memória do *Hadoop cluster* implantado na nuvem privada, em *GB*.

Os tempos de mapeamento(T_{map}) e de redução (T_{red}) podem ser representados por distribuições de probabilidade diferentes da distribuição de probabilidade exponencial. Neste caso, a média e o desvio-padrão de T_{map} e T_{red} são calculados para analisar a distribuição de probabilidade que melhor representa esses tempos, através da técnica de aproximação por fases.

5.2 Modelos de custo de aplicações *Big data* em nuvens privadas

Em ambientes reais, diversos custos podem ser considerados para a implantação de ambientes de nuvem computacional para atender requisições geradas por transações *Big Data*. Três principais componentes de custo são definidos (PALIAN, 2018):

- Custo por unidade de *Rack* - representa os custos relacionados ao *hardware*, manutenção de infraestrutura de nuvem e custos do espaço alocado para a implantação

da nuvem. O custo de *hardware* os custos de componentes de servidores. O custo de manutenção da infraestrutura denota ferramentas de segurança(como *firewalls*), *switches* e outros equipamentos pertinentes ao ambiente de nuvem. Por fim, o cust de pessoal ou *staff*, representa a equipe para gerenciar, monitorar e resolver problemas na infraestrutura de nuvem para garantir disponibilidade e confiabilidade.

- Custo por *GB* de memória virtual - representa a quantidade de memória RAM disponível para os usuários da nuvem computacional e a sua aquisição.
- Custo por *GB* de disco virtual oferecido aos usuários - calculado pelos custos associados à operacionalidade de dispositivos de armazenamento virtual da nuvem e pela aquisição de novas tecnologias de armazenamento.
- Custos compartilhados - custo relacionados ao consumo de energia e refrigeração da infraestrutura de nuvem, que podem ser aplicados aos usuários ou não, no caso de nuvens públicas.

Os modelos de custo tem como objetivo estimar o custo de aplicações *Big Data* em ambientes de nuvem privada e consideram alguns dos componentes acima citados, devido à sua aplicabilidade. Estes modelos consideram dentre os diversos custos que podem ser associados, o custo de infraestrutura de nuvem, custo de consumo de energia e o custo de aquisição de *software*. Neste caso, o custo total da implantação da aplicação *Big Data* na nuvem privada é composto pelo custo da infraestrutura de nuvem privada, pelo custo da energia elétrica que a aplicação *Big Data* demanda em seu tempo de execução e pelo custo de *software* incorrido com a aquisição de programas pertinentes ao ambiente *Big Data* e a plataforma de nuvem privada.

O custo da infraestrutura de nuvem privada está relacionado ao custo das máquinas físicas, *switches*, roteadores e *nobreaks*, necessários para a configuração do ambiente *Big Data* na nuvem privada. O custo da energia elétrica está relacionado à energia elétrica consumida durante a aplicação das cargas de trabalho no ambiente *Big Data* na nuvem privada. O custo de *software* refere-se ao custo de aquisição dos programas necessários para a implantação de ambientes *Big Data* na nuvem privada. O custo total da aplicação *Big Data* executada em nuvem privada é definido pela Equação 5.1, os componentes de custo(C_{infra}, $C_{energia}$ e $C_{software}$) representam o modelo de custo de infraestrutura do ambiente *Big Data* na nuvem privada, modelo de custo de energia elétrica e o modelo de

custo de *software*.

$$C_{total} = C_{infra} + C_{energia} + C_{software} \tag{5.1}$$

- **O modelo de custo de infraestrutura do ambiente *Big Data* na nuvem privada(C_{infra})** - representa os custos associados aos equipamentos da nuvem privada como quantidade de computadores, roteadores, *switches*, e *nobreaks*. A Equação 5.2 denota o custo de infraestrutura do ambiente *Big Data* na nuvem privada.

$$C_{infra} = \sum_{i=1}^{n} Nhost_i \times Chost_i + \sum_{i=1}^{n} Ns_i \times Cs_i + \sum_{i=1}^{n} Nr_i \times Cr_i + \sum_{i=1}^{n} Nn_i \times Cn_i \tag{5.2}$$

  Em que: n indica os tipos de computadores, *switches*, roteadores e *nobreaks* na nuvem privada;

  $Nhost_i$ representa o número de computadores do mesmo modelo que compõem a nuvem privada;

  $Chost_i$ representa o custo unitário de cada computador do mesmo tidpo na nuvem privada;

  Ns_i = número de *switches* do mesmo tipo na nuvem privada;

  Cs_i = custo unitário de cada *switch* do mesmo tipo na nuvem privada;

  Nr_i = número de roteadores do mesmo tipo que compõem a nuvem privada;

  Cr_i = custo unitário de cada roteador do mesmo tipo na nuvem privada;

  Nn_i = quantidade de *nobreaks* do mesmo modelo na nuvem privada;

  Cn_i = custo unitário de cada *nobreak* do mesmo modelo que compõe a nuvem privada.

- **Modelo de custo de energia elétrica($C_{energia}$)** - representa os custos associados ao consumo de energia elétrica durante a execução da aplicação *Big Data* no ambiente de nuvem privada. Considera a potência consumida em kilowatts(ΔkW) demandada durante o seu tempo de execução(ΔT). O parâmetro δ é uma constante que indica o valor monetário do kWh. A Equação 5.3 denota o custo de energia elétrica, no qual i representa o número de aplicações *Big data* submetidas para a infraestrutura

de nuvem privada no somatório e n o número total de aplicações implantadas.

$$C_{energia} = \sum_{i=1}^{n} \Delta kW_i \times \Delta T_i \times \delta \tag{5.3}$$

- Modelo de custo de *software*($C_{software}$) - denotado por todos os custos associados a programas adotados para a implantação de aplicações *Big data* em nuvens privadas. É composto de custo de sistemas operacionais, custo da plataforma de nuvem privada, e custo da aplicação *Big data*. É considerado apenas uma plataforma de nuvem privada adotada e aquisição de uma aplicação *Big data* (Equação 5.4).

$$C_{software} = \sum_{i=1}^{n} Nsist_i \times Csist_i + Cap_i + (plat \times Cplat) \tag{5.4}$$

Em que:
n = tipos de aplicação *Big Data* na nuvem privada;
$Nsist_i$ = número de sistemas operacionais;
$Csist_i$ = custo do sistema operacional;
Cap_i = custo unitário da aplicação *Big Data*;
$plat$ = plataforma de nuvem privada adotada;
$Cplat$ = custo unitário da plataforma de nuvem privada.

5.3 Considerações Finais

Este capítulo apresentou os modelos de desempenho e custo adotados neste trabalho para avaliar aspectos de desempenho e custo de ambientes *Big Data* implantados em ambientes de nuvem privada. O modelo de desempenho é baseado em redes de Petri estocásticas, no qual há a representação de cargas de trabalho sendo enviadas para o ambiente *Big Data* na nuvem privada, que analisa o *data set* e devolve para o usuário a informação resultantes da análise. Os modelos de custo contemplam aspectos como o custo da infraestrutura do ambiente *Big Data* na nuvem privada, que oferece os recursos computacionais para a análise do conjunto de dados, custo de energia elétrica e custo

com programas necessários para a execução das aplicações *Big Data*. O seguinte capítulo apresenta o estudo de caso demonstrando a aplicabilidade da metodologia, do modelo de desempenho e dos modelos de custo em um ambiente real.

6 Estudo de caso

Este capítulo apresenta o estudo de caso realizado para demonstrar a aplicabilidade da metodologia proposta em um cenário real de análise Big Data, em que uma infraestrutura de nuvem privada provisiona os recursos computacionais para as transações *Big Data*. Neste estudo de caso será avaliado o desempenho e custo do *Hadoop cluster* na nuvem privada. Além disso, esse estudo de caso apresenta a análise da capacidade máxima de carga de trabalho suportada pelos *data nodes* instanciados no ambiente de nuvem privada, considerando as métricas de utilização de processador e utilização de memória.

6.1 Introdução

O cenário real avaliado no estudo de caso é um ambiente *Big Data* configurado na nuvem privada. Este ambiente *Big Data* é formado pelo *Hadoop clusters*, no qual o mecanismo *MapReduce* (DEAN; GHEMAWAT, 2010) é adotado para analisar os *data sets* e assim, gerar a carga de trabalho. A plataforma de nuvem privada considerada para gerenciar todos os processos de instanciação de máquinas virtuais, adição de ofertas de serviço é o Apache Cloudstack (CLOUDSTACK, 2018).

6.2 Estudo de Caso: Avaliação de Desempenho e Custo de Aplicações *Big Data* na Nuvem Privada

Este estudo de caso tem como objetivo avaliar o desempenho e o custo de ambientes *Big Data* suportados por infraestruturas de nuvem privada. Fatores impactantes em diferentes variáveis de resposta, como tempos de execução, utilização de processador, utilização de memória, consumo de energia e custo de *data nodes* configurados em máquinas virtuais, são identificados. Por outro lado, o modelo de desempenho validado é adotado para estimar utilização de processador e memória dos *data nodes* configurados na nuvem privada e os modelos de custo avaliam o custo de infraestrutura, custo associado ao consumo de energia elétrica, e custo de *software* de ambientes *Big Data* configurados em nuvens privadas.

6.2.1 Entendimento e configuração do ambiente de anáise *Big Data* em nuvem privada

A nuvem privada foi configurada com 8 máquinas físicas, sendo que uma máquina física é o gerenciador da nuvem e as outras 7 máquinas físicas são nós que provisionam a capacidade computacional para instanciar as máquinas virtuais. Um roteador isola as máquinas físicas da nuvem privada e um switch com 24 portas foi utilizado para conectar todas as máquinas físicas entre si. O *switch* comporta até 24 *hosts físicos* e um *no-break* de capacidade de 1500 VA e suporta até 8 *hosts* físicos. A Tabela 4 mostra as configurações das máquinas físicas da nuvem privada implantada.

Tabela 4 – Características dos *hosts* da nuvem privada

Recurso	**Carcertísticas**
Processador	*Intel core i5*
Memória	8,00 *GB*
Disco	1 *TB*
Sistema operacional	CentOS 7 minimal 64 *bits*
Hypervisor	*KVM*

O sistema operacional para as máquinas físicas assim como as máquinas virtuais instanciadas na nuvem privada foi o CentOS em sua versão mínima (CENTOS, 2018), sem interface gráfica para reduzir *overheads* que podem ser causados pelo consumo de recursos por parte do sistema operacional. Esta escolha foi feita considerando o sistema operacional com menos requisitos de *hardware* para a sua instalação e recomendado para ambientes de negócios devido a sua estabilidade e aspectos de segurança (LONGEN, 2018).

A Figura 14 denota o ambiente *Big Data* configurado na nuvem privada. Esta figura apresenta o ambiente adotado para execução dos experimentos em que as máquinas virtuais são instanciadas para formar o *Hadoop cluster* que irá analisar os conjuntos de dados através de mecanismos distribuídos. Na figura, a infraestrutura de nuvem privada é representada com suas máquinas físicas conectadas a um *switch* e roteador. Máquinas virtuais são instanciadas neste ambiente de nuvem privada com o sistema operacional CentOS (CENTOS, 2018). Com as máquinas virtuais instanciadas, um *master node* e *data nodes* são configurados nas máquinas virtuais na nuvem privada.

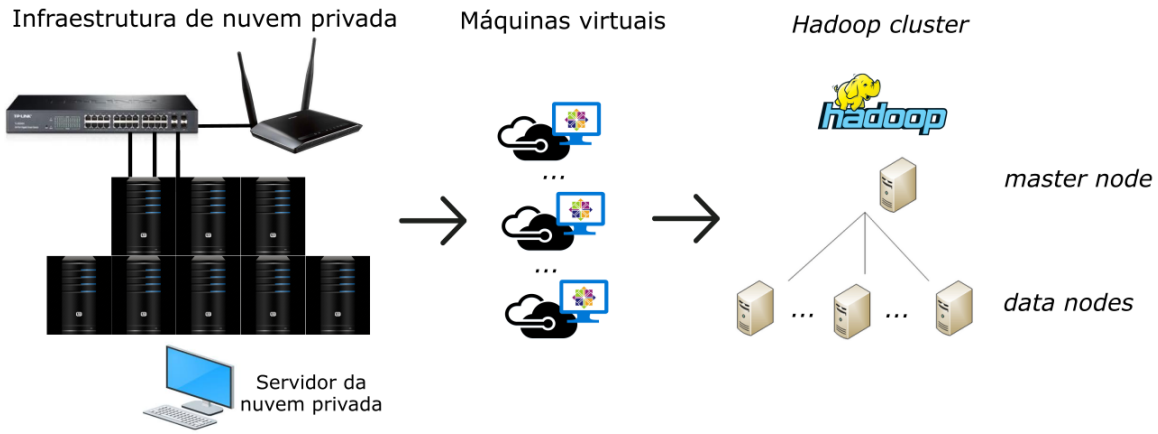

Figura 14 – Ambiente *Big Data* implantado na nuvem privada (FONTE: Próprio Autor)

6.2.2 Seleção de métricas de desempenho e consumo de energia em amibentes *Big Data*

As seguintes métricas adotadas neste trabalho foram: Tempo de execução(segundos), que consiste no tempo demandado para o atendimento das cargas de trabalho geradas pela análise *Big Data*; Utilização de processador(%) que representa a porcentagem média em que os processadores das máquinas virtuais estão ocupadas atendendo as cargas de trabalho geradas; Utilização de memória(%) que representa a quantidade média de memória utilizada durante a execução dos experimentos; Consumo de energia elétrica(*Watts*), que denota a quantidade de *Watts* consumidos durante a execução dos experimentos; Custo de infraestrutura que representa o custo dos equipamentos da nuvem privada e Custo de *software* que representa o Custo de aquisição de programas e executáveis. O desempenho do cluster Hadoop ao analisar datasets depende da capacidade de processamento e de memória principal disponível (ALAPATI, 2016).

6.2.3 Planejamento de Experimentos para a aplicação *Big Data* em infraestruturas de nuvens privadas

A oferta de serviço da nuvem privada, o tamanho do *cluster* e o tamanho do *dataset*, são considerados como fatores representativos para o desempenho de análise *Big Data* em ambientes de nuvem privada (ECKROTH, 2016; ZHANG; SAKR, 2014; ADHIKARI et al., 2017). Nesse trabalho, o planejamento de experimentos adotou esses fatores com diferentes níveis. A oferta de serviço define a capacidade de provisionamento da nuvem privada

para os *data nodes* configurados em máquinas virtuais. A flexibilidade da nuvem permite que as mais variadas ofertas de serviços possam ser definidas com várias capacidades de processamento e de memória. A Tabela 5 denota as ofertas de serviço adotadas para o planejamento de experimentos.

Tabela 5 – Ofertas de serviço da nuvem privada

Oferta de serviço	**Configuração**
Small	CPU: 2 Cores; Memória: 4 *GB*; HD: 100*GB*
Medium	CPU: 4 Cores; Memória: 6 *GB*; HD: 100*GB*
Large	CPU: 8 Cores; Memória: 8 *GB*; HD: 100*GB*

O segundo fator a ser abordado é o número de *data nodes* para analisar os conjuntos de dados. Representa a escalabilidade da aplicação *Big Data* quando mais máquinas virtuais são instanciadas na infraestrutura de nuvem. Com a restrição de capacidade da nuvem privada adotada neste estudo de caso, os níveis estabelecidos para o fator Número de *data nodes* foram 3, 5 e 7 *data nodes*.

O terceiro e último fator é o tamanho do *dataset*. Analisar o impacto deste fator é essencial em um ambiente de *Big Data*, visto que diferentes volumes de dados podem ser submetidos para análise na nuvem privada. A Tabela 6 apresenta os fatores e os níveis adotados para os experimentos.

Tabela 6 – Fatores e níveis do planejamento de experimentos

Fatores	**Níveis**		
Ofertas de serviço	*Small*	*Medium*	*Large*
Número de *datanodes*	3	5	7
Intensidade da carga de trabalho	4 *GB*	7 *GB*	10 *GB*

Os fatores e níveis da tabela anterior foram adotados de acordo com a capacidade de processamento e memória da nuvem privada. Esta atividade definiu 3 fatores que foram variados em 3 níveis, gerando 27 experimentos através de um planejamento de experimentos fatorial (MONTGOMERY, 2017). Foram consideradas 20 replicações por experimento que proporcionaram uma média e um desvio-padrão das métricas de desempenho. A Tabela 7 apresenta os cenários gerados pelo planejamento de experimentos fatorial.

Tabela 7 – Cenários gerados pelo planejamento de experimentos fatorias

Cenário	**Oferta de serviço**	**Número de data nodes**	**Tamanho do *dataset*(GB)**
1	*Small*	3	4
2	*Small*	3	7
3	*Small*	3	10
4	*Medium*	3	4
5	*Medium*	3	7
6	*Medium*	3	10
7	*Large*	3	4
8	*Large*	3	7
9	*Large*	3	10
10	*Small*	5	4
11	*Small*	5	7
12	*Small*	5	10
13	*Medium*	5	4
14	*Medium*	5	7
15	*Medium*	5	10
16	*Large*	5	4
17	*Large*	5	7
18	*Large*	5	10
19	*Small*	7	4
20	*Small*	7	7
21	*Small*	7	10
22	*Medium*	7	4
23	*Medium*	7	7
24	*Medium*	7	10
25	*Large*	7	4
26	*Large*	7	7
27	*Large*	7	10

6.2.4 Geração da carga de trabalho *Big Data*

A carga de trabalho é caracterizada pelo tamanho do *data set* que o usuário deseja analisar no ambiente. Diferentes tipos de conjuntos de dados, cada um em seu ambiente de estudo, como dados de redes sociais, dados de saúde e dados meteorológicos (ASSUNÇÃO et al., 2015; HASHEM et al., 2015; YANG et al., 2017).

Para determinar o tipo de *data set* a ser adotado para os experimentos, uma análise prévia foi realizada para identificar quais temas são mais citados em artigos relacionados à *Big Data* e *Cloud computing* em bases científicas tais como IEEE Xplore, Springer, Elsevier e ACM. Foi constatado que os temas mais citados estão relacionados à análise de dados de redes sociais e de dados de saúde. Esta análise considerou as citações dos termos *Big Data* e *Cloud computing* em 750 artigos. Uma vez que o tema *Social network* apresentou a maior citação considerando *Big Data* e *Cloud computing*, o processo de captura dos dados foi iniciado considerando a rede social *Twitter*. Esta rede social propicia informações através da análise de sentimento de diversas áreas, como a política (MALIK et al., 2018).

Uma vez definido o tipo de *data set* a ser analisado no ambiente *Big data* suportado pela nuvem privada, a carga de trabalho é caracterizada por conjuntos de *posts* relativos aos comentários dos usuário da rede social *Twitter*, em que o objetivo é avaliar a posição política destes em relação às eleições presidenciais de 2018. O *data set* composto de comentários da rede social foram capturados através do algoritmo apresentado no Apêndice A, através da ferramenta *R software* (R, 2018).

Em 2018 foi realizada a eleição para presidência do Brasil e o *Twitter* apresentou diversas postagens de eleitores brasileiros sobre suas posições políticas. O *data set* criado para os estudos de caso foram baseados nos *posts* dos eleitores brasileiros no Twitter, a partir de 28 de setembro em 2018 até 28 de outubro de 2018, período entre o primeiro e segundo turno das eleições para presidente do Brasil. *Hashtags* relacionados aos candidatos como #Bolsonaro2018, #Bolsonaro2018, #BrasilViraHaddad e #Haddad2018, dentre outros, foram adotadas para identificação das opiniões dos eleitores brasileiros e capturar os conjuntos de dados formados por uma grande quantidade de *posts* da rede social.

6.2.5 Execução das aplicações *Big Data* na nuvem privada

Inicialmente, o *data set* foi armazenado pelo *master node* no *HDFS* e em seguida os campos de interesse do conjunto de dados foram selecionados, de acordo com os objetivos da análise dos dados. Os campos de interesse neste caso são o usuário e o campo que contém os *posts* do *Twitter* no *data set*. O *bag* de palavras corresponde às tuplas(pares chave-valor) formadas pelas palavras contidas no *dataset* (GATES; DAI, 2016). A partir do *bag* de palavras, é possível agrupar as mesmas palavras e ordená-las de forma a contabilizar o número de vezes que cada uma foram citadas no *data set*. Por fim, os resultados são armazenados no *HDFS* para posterior análise das informações geradas.

O processo para capturar os dados do *Twitter* está no Apêndice A no final deste documento e o cálculo das palavras mais citadas na rede social no Apêndice B. A Figura 15 denota como ocorreu este processo.

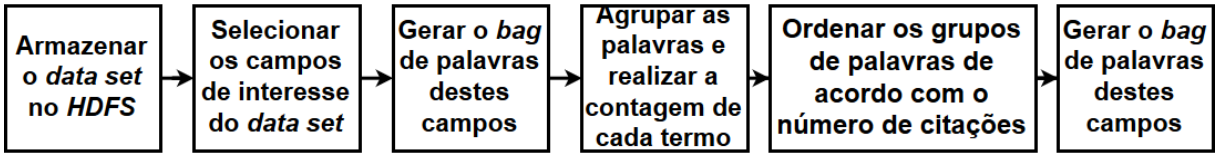

Figura 15 – Processo de análise do *dataset* no *Hadoop cluster*(FONTE: Próprio Autor)

6.2.6 Medição de desempenho e consumo de energia da aplicação *Big Data* na nuvem privada

A medição de desempenho consiste da coleta dos tempos de execução, das utilizações médias de processadores e de memórias e consumo de energia dos *data nodes* configurados nas máquinas virtuais da nuvem privada, durante as aplicações das cargas de trabalho.

As ferramentas adotadas para medição de utilização de processador e de memória dos *data nodes* configurados em máquinas virtuais da nuvem privada foram o *MPSTAT* e o *VMSTAT* do pacote *SYSSTAT* (JUVE et al., 2015). A medição do consumo de energia foi auxiliada pela plataforma *online* Arduino Create (ARDUINO, 2018). O script para a coleta de consumo de energia da aplicação *Big Data* na nuvem privada está no Apêndice E deste trabalho. A medição do consumo de energia foi realizada através de um dispositivo *IoT*, que mede esta métrica por indução no *nobreak* (SILVA, 2017), como mostra a Figura 16. Nesta figura, as máquinas físicas da infraestrutura de nuvem privada são ligadas a um

mesmo *nobreak* e na saída do *nobreak*, o sensor de coleta de consumo de energia elétrica é instalado, sem ser invasivo nas medições. O dispositivo Arduino fornece *logs* relativos ao consumo de energia a cada 1 segundo durante o tempo de execução de cada experimento.

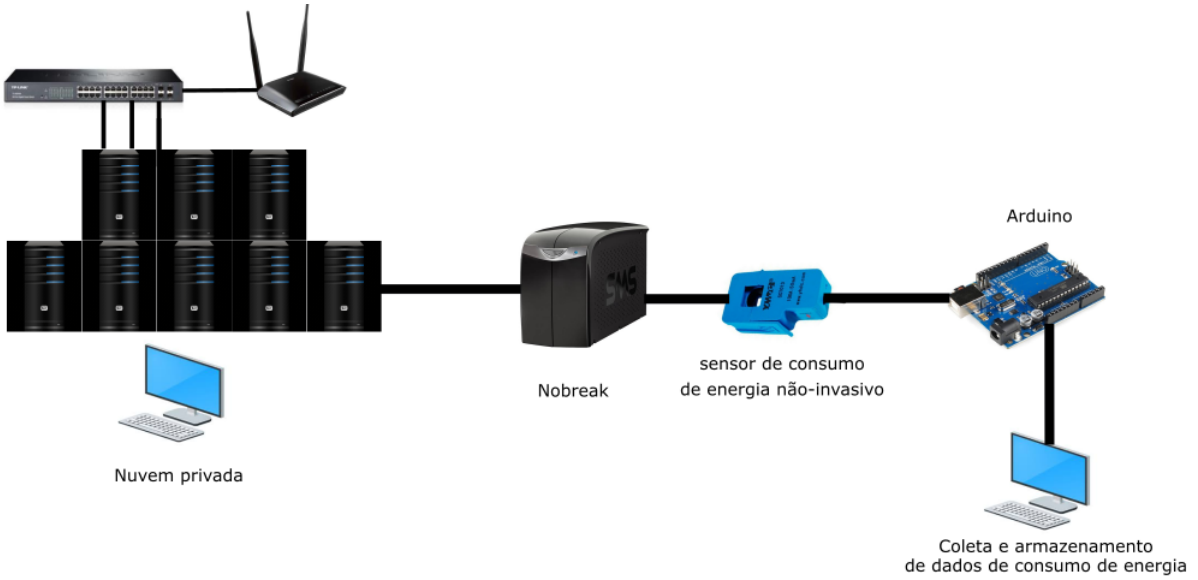

Figura 16 – Medição de consumo de energia nos experimentos (FONTE: Próprio Autor)

O intervalo de amostragem adotado para a coleta de tempos de execução, utilizações médias de processamento, memória e consumo de energia é de 1 segundos, com um intervalo de *warm-up* de 1 minuto e um intervalo de *cooldown*, após o processamento do *data set* de 1 minuto. Com estes tempos, é possível avaliar o momento que a carga de trabalho inicia o consumo de recursos assim como avaliar o momento que a carga de trabalho é finalizada, liberando os recursos. O *script* de monitoramento captura a utilização dos recursos durante os experimentos. *Scripts* de monitoramento de processador(Apêndice B) e de memória(Apêndice C) em *Shell script* (TANENBAUM, 2009) assim como de captura de consumo de energia em tempo real foram desenvolvidos para a medição de desempenho. Todos os dados da medição de desempenho e custo foram armazenados em *logs*(.csv) para posterior análise estatística, avaliando as médias e os desvios-padrões para cada cenário gerado no planejamento de experimentos. As máquinas foram isoladas, evitando a execução de programas que não faziam parte dos experimentos e ocorreram reinicializações após a execução de cada replicação dos experimentos.

6.2.7 Análise estatística das métricas de desempenho e de consumo de energia

Para cada experimento, foram coletadas as utilizações médias de processador e de memória, o consumo de energia elétrica e os tempos de execução demandados. A retirada de *outliers* envolveu a retirada de valores que podem ter sido causados por erros menores tais como perturbações no ambiente de medição. Estes *outliers* são pontos fora do comportamento normal e são denidos de acordo com quartis da série de dados das métricas de desempenho (GUPTA; GUTTMAN, 2014). Foram geradas médias e desvios-padrões referentes à cada cenário configurado de acordo com os fatores e níveis definidos no planejamento de experimentos. Após a retirada destes pontos, as médias e os desvios-padrões são calculados representando as métricas de desempenho e consumo de energia em cada experimento.

Para realizar a análise estatística dos experimentos, o Minitab *software* estatístico (MINITAB, 2018) foi adotado, no qual se avaliou o impacto dos fatores considerados(oferta de serviço, número de *data nodes* e tamanho do *data set*) nas variáveis de resposta (Tempo de execução, Utilização de Processador e de Memória, e Consumo de energia elétrica e custo).

Um maior impacto pode ser notado nos gráficos, quando as retas que ligam os pontos possuem uma maior ou menor inclinação. Este impacto é caracterizado pela capacidade de aumentar ou diminuir o valor das métricas de utilizações de recursos e consumo de energia médios, ou variáveis de resposta.

As Figuras 17, 18, 19 e 20, apresentam, respectivamente, o impacto das variáveis de entrada(oferta de serviço, número de *data nodes* e tamanho do *data set*) nas variáveis de resposta(tempo de execução, utilização de processador e de memória, consumo de energia elétrica dos *data nodes* configurados na nuvem privada). Cada ponto do gráfico denota o valor médio para cada nível dos fatores(oferta de serviço, número de *data nodes* e tamanho do *data set*) (MONTGOMERY, 2017).

A Figura 17 apresenta as médias de tempos de execução considerando os fatores e níveis adotados no planejamento de experimentos.

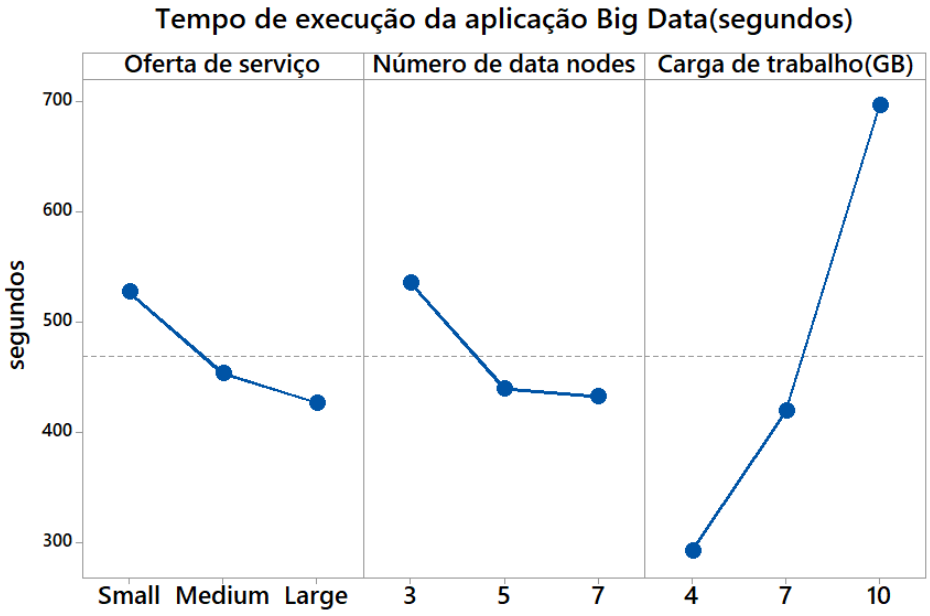

Figura 17 – Tempos médios de execução da aplicação *Big Data* na nuvem privada (FONTE: Próprio autor)

A carga de trabalho impactou de forma mais significativa no tempo de execução da aplicação *Big Data* em comparação com a oferta de serviço e a quantidade de *data nodes* no *cluster*. O tempo de execução depende principalmente do tamanho de *data set* que o usuário deseja processar na nuvem privada. Aspectos de gerenciamento da nuvem privada como a oferta de serviço e a quantidade de *data nodes* configurados em máquinas virtuais da nuvem privada, não apresentaram impacto significativo nesta métrica de desempenho.

A Figura 18 apresenta as utilizações médias de processador de *data nodes* instanciados através de máquinas virtuais da nuvem privada para a análise *Big Data*.

A oferta de serviço apresentou um impacto significativo nesta métrica e demonstrou ser o fator mais importante para ambientes em que se deseja reduzir a utilização de processamento. Nestes ambientes, a quantidade de máquinas virtuais e o tamanho do conjunto de dados da rede social não impactaram a utilização de processadores.

A Figura 19 denota as médias das métricas de utilização de memória dos *data nodes* configurados em máquinas virtuais da nuvem privada. A oferta de serviço apresentou um impacto significativo nesta métrica, como no caso anterior(utilização de processador). Porém, os outros dois fatores também apresenta impacto significativo. Neste caso, a quantidade de máquinas virtuais instanciadas para analisar o conjunto de dados da rede social diminuiu a utilização de memória, mas não impactou na redução da utilização de processador. Isto se deve ao fato que a memória possui um impacto maior no desempenho

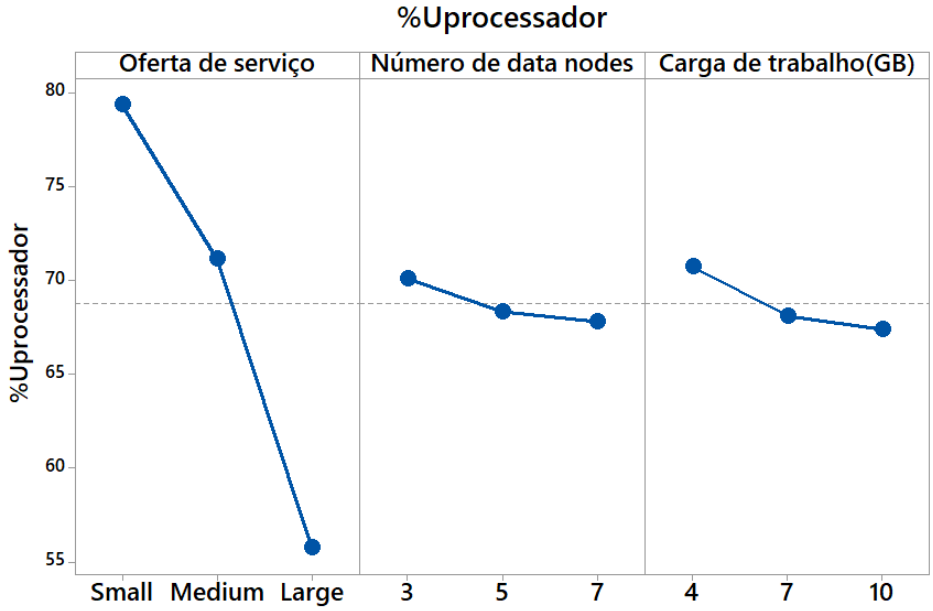

Figura 18 – Utilização média de processador dos *data nodes* instanciados na nuvem privada (FONTE: Próprio autor)

de *Hadoop clusters* (ALAPATI, 2016). Para reduzir a utilização de memória dos *data nodes*, seria necessário aumentar a oferta de serviço, aumentar a quantidade de *data nodes* no *Hadoop cluster*.

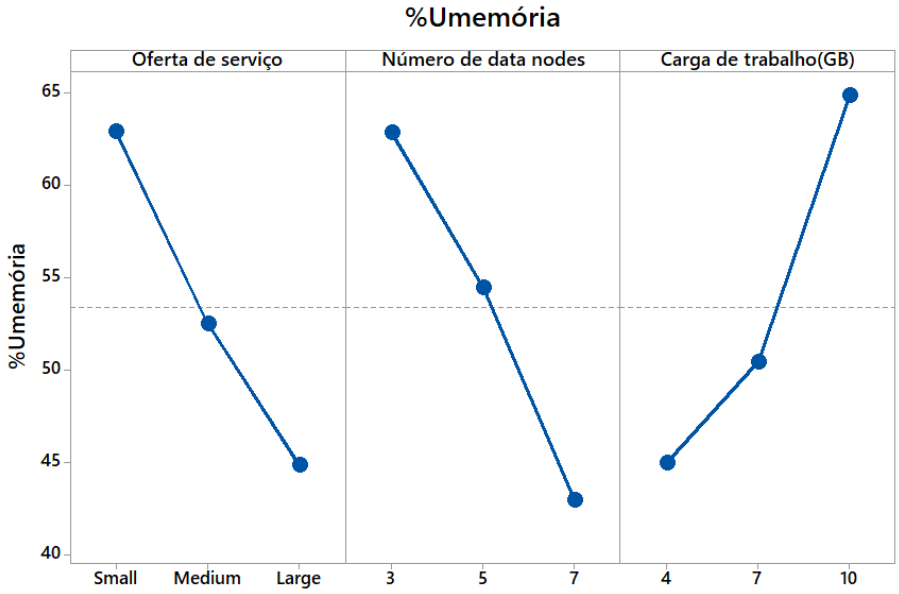

Figura 19 – Utilização média de memória das máquinas virtuais na nuvem privada (FONTE: Próprio autor)

O consumo médio de energia elétrica do ambiente *Big Data* configurado na nuvem privada considerando os diferentes fatores e seus níveis é apresentado na Figura 20.

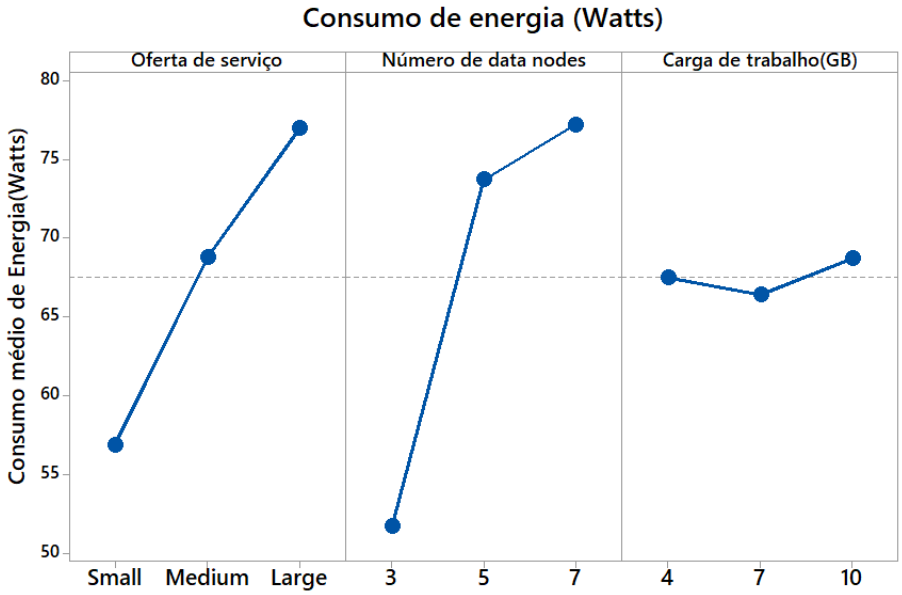

Figura 20 – Consumo de energia médio do ambiente *Big Data* implantado infraestrutura de nuvem privada (FONTE: Próprio autor)

A Figura 20 mostra que a oferta de serviço e a quantidade de *datanodes* no *cluster* apresentaram um alto impacto nesta métrica. O fator tamanho da carga de trabalho permaneceu quase constante com a variação de seus níveis. Isto sugere que o consumo de energia destes ambientes, são principalmente impactadas pela quantidade de máquinas virtuais instanciadas, seguido pela oferta de serviço.

6.2.8 Geração dos modelos de custo da aplicação *Big Data* na nuvem privada

Os modelos de custo adotados para avaliar as aplicações *Big Data* na nuvem privada calculam o custo com infraestrutura (Equação 5.2), custo com consumo de energia elétrica (Equação 5.3) e custo com *software*(Equação 5.4). A Tabela 8 apresenta os custos de infraestrutura considerando os equipamentos que compõem a infraestrutura da nuvem privada. O custo da infraestrutura do *Hadoop cluster* configurado na nuvem privada calculado através da Equação 5.2 foi de **R$ 16.773,77**.

O cálculo do custo de energia elétrica da aplicação *Big Data* na nuvem privada foi realizado através da Equação 5.3. A Tabela 9 apresenta o consumo de energia elétrica das máquinas físicas da nuvem privada para cada cenário gerado pelo planejamento de

Tabela 8 – Custo dos equipamentos que compõem a infraestrutura da nuvem privada

Equipamentos	**Quantidade**	**Valor unitário(R$)**	**Valor total(R$)**
Computador i5-5200U 8GB 1TB	8	1.905,87	15.246,96
Switch 16 portas	1	411,75	411,75
Nobreak 1500 VA	1	1.047,06	1.047,06
Roteador	1	68,00	68,00

experimentos, convertidos kW e considerando um δ de 0,66642619, valor em reais do kiloWatt-hora definido pela companhia elétrica local (CELPE, 2018). O custo de energia elétrica para cada cenário é apresentado na Tabela 9. O termo kW indica o consumo médio de energia do cenário, em kiloWatts. O tempo médio demandado para cada Cenário é convertido em Horas(Tempo(Horas)). O Custo de consumo de energia(**Energia(R$)**) corresponde ao custo associado à demanda de energia elétrica para cada configuração do ambiente *Big Data*. O custo de *software* é R$ 0,00, devido ao fato de que foram utilizados apenas programas *open source* para a execução aplicação *Big Data* na nuvem privada. A última coluna denota o custo total para cada cenário, considerando a Equação 5.1 do Capítulo 5.

A Figura 21 denota os fatores mais impactantes para o custo da aplicação através do mesmo planejamento de experimentos adotado no Minitab (MINITAB, 2018) para as métricas de desempenho, considerando a variável de resposta como sendo o Custo(R$). Os fatores Oferta de serviço e Número de *data nodes* apresentaram um maior impacto na variação do custo do ambiente *Big Data* na nuvem privada.

A Tabela 10 apresenta as métricas ou variáveis de resposta condicionadas à seu(s) fatore(s) mais impactantes. Os fatores impactantes foram identificados através de um planejamento de experimentos fatorial. Esta tabela contribui para identificar possíveis gargalos do ambiente e apresentando fatores de maior impacto para o ambiente *Big Data* suportado pela infraestrutura de nuvem privada.

Tabela 9 – Custo de energia elétrica para cada cenário gerado pelo planejamento de experimentos

Cenário	***kW***	**Tempo(Horas)**	**Valor do kWh(R$)**	**Energia(R$)**	**Custo Total(R$)**
1	0,213	0,107	0,666	0,015	16.773,78
2	0,214	0,139	0,666	0,019	16.773,80
3	0,233	0,254	0,666	0,039	16.774,02
4	0,238	0,065	0,666	0,010	16.773,78
5	0,245	0,097	0,666	0,015	16.773,78
6	0,238	0,168	0,666	0,026	16.773,89
7	0,259	0,076	0,666	0,013	16.773,78
8	0,241	0,119	0,666	0,019	16.773,79
9	0,250	0,235	0,666	0,039	16.773,99
10	0,238	0,091	0,666	0,014	16.773,78
11	0,222	0,130	0,666	0,019	16.773,79
12	0,244	0,207	0,666	0,033	16.773,93
13	0,262	0,065	0,666	0,011	16.773,78
14	0,256	0,087	0,666	0,014	16.773,78
15	0,246	0,175	0,666	0,028	16.773,80
16	0,262	0,066	0,666	0,011	16.773,78
17	0,258	0,096	0,666	0,016	16.773,79
18	0,251	0,172	0,666	0,028	16.773,80
19	0,236	0,084	0,666	0,013	16.773,78
20	0,239	0,128	0,666	0,020	16.773,79
21	0,247	0,197	0,666	0,032	16.773,80
22	0,262	0,070	0,666	0,012	16.773,78
23	0,273	0,097	0,666	0,017	16.773,79
24	0,262	0,142	0,666	0,024	16.773,79
25	0,257	0,077	0,666	0,013	16.773,78
26	0,269	0,098	0,666	0,017	16.773,78
27	0,267	0,144	0,666	0,025	16.773,80

6.2.9 Geração do modelo de desempenho

O modelo de desempenho adotado para avaliar o desempenho de aplicações *Big Data* em infraestruturas de nuvem privada é apresentado na Seção 5.1, do Capítulo 5.

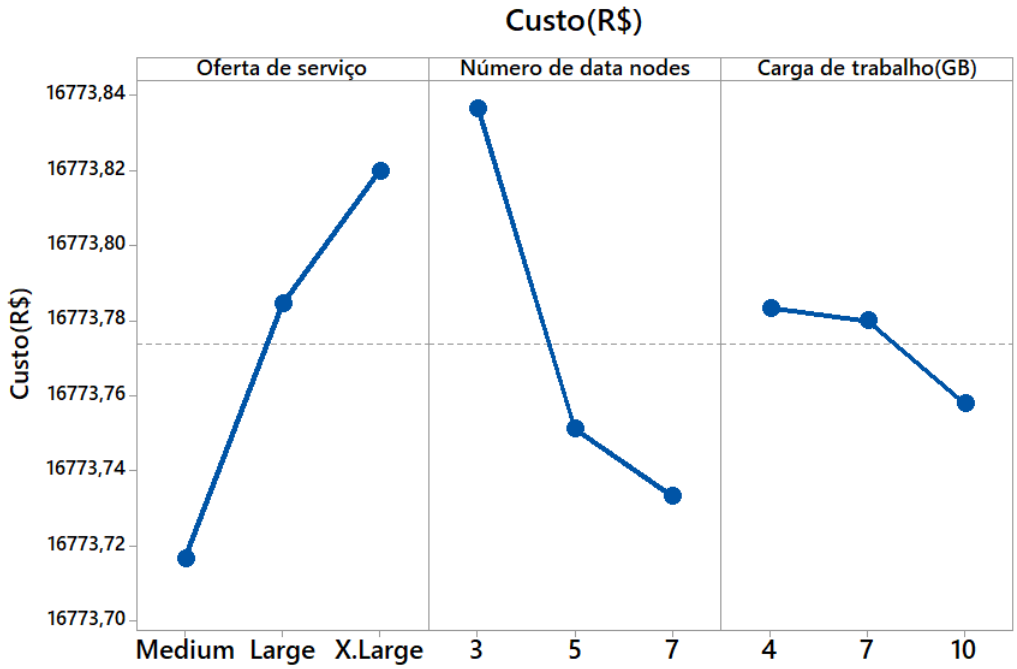

Figura 21 – Custo médio do ambiente *Big Data* na nuvem privada (FONTE: Próprio autor)

Tabela 10 – Análise de sensibilidade de métricas de desempenho - Fatores impactantes para o desempenho de aplicações *Big Data* na nuvem privada

Variável de resposta	**Fatores mais impactantes**
Tempo de execução(segundos)	Tamanho *data set*
Utilização de processador(%)	Oferta de serviço da nuvem privada
Utilização de memória(%)	Oferta de serviço/Quant. de *data nodes*/Tamanho *data set*
Consumo de energia elétrica(*Watts*)	Oferta de serviço/Quant. de *data nodes*
Custo(R$)	Oferta de serviço/Quant. de *data nodes*

6.2.10 Avaliação das propriedades do modelo de desempenho

A ferramenta HiPS (HIPS, 2018) realizou a análise das propriedades comportamentais e estruturais do modelo de desempenho proposto(Seção 5.1, Capítulo 5) e concluiu que não há *deadlocks* assim como todas as transições são disparadas durante a execução do modelo e que todos os lugares foram acessados por marcações. Foram analisadas também as propriedades de alcançabilidade e *liveness* e foi verificado que estas propriedades são satisfeitas no modelo.

6.2.11 Refinamento do modelo de desempenho

Os tempos de mapeamento e de redução foram adotados para o refinamento do modelo de desempenho proposto. Esses tempos foram medidos em 20 replicações em cada um dos 27 experimentos planejados conforme a Seção 6.1.1(Ver Tabela 7).

De acordo com (JAIN, 1991), o pesquisador pode adotar uma parte do problema geral para gerar, refinar, e validar o modelo de desempenho e verificar este modelo para outros cenários. Neste contexto, os cenários considerados como parte do problema foram os de 3 *data nodes* no *Hadoop cluster*(Cenários 1 até 9). Em outro momento, este modelo é verificado para os cenários com 5 e 7 *data nodes*(Cenários 10 até 18) (Ver Tabela 7).

Testes Chi-quadrado foram realizados para determinar a distribuição de probabilidade que representa os comportamentos dos tempos de mapeamento e de redução dos *data sets*, medidos nos cenários com 3 data nodes(Cenários 1 até 9). Os tempos médios de mapeamento para os Cenários de 3 *data nodes*(Cenários 1 até 9) foram analisados e a distribuição de probabilidade Beta apresentou o menor erro quadrático médio (0,38%) para representação desse tempo. O tempo médio de redução para os Cenários com 3 *data nodes* podem ser representados pela distribuição de probabilidade Normal, com erro quadrático médio de 4%.

Visto que os tempos de mapeamento e de redução de não são representados por uma distribuição de probabilidade exponencial, logo, é necessário identicar que distribuição expolinomial representa esses tempos (MACIEL et al., 2017). 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*(ver Tabela 7) foram calculados para identificar a distribuição expolinomial que melhor representa o comportamento desses tempos.

A Tabela 11 denota a distribuição de probabilidade que melhor representa os tempos de mapeamento 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 12 apresenta os parâmetros da distribuição Hipoexponencial para as transições **Mapeamento do *dataset*** e **Redução do *dataset*** do modelo de desempenho.

A Figura 22 denota o modelo de desempenho refinado considerando a identificação da distribuição de probabilidade que melhor representa os tempos de mapeamento e redução do *data set* e os parâmetros da Tabela 12.

Tabela 11 – Média, Desvio-Padrão e Distribuição de Probabilidade Expolinomial

Tempos	**Tempo médio(segundos)**	**Desvio-padrão**	$1/C_v$	**Dist. Prob.**
Mapeamento	457,78	192,50	2,32	Hipoexponencial
Redução	61,82	20,48	3,02	Hipoexponencial

Tabela 12 – Parâmetros da distribuição Hipoexponencial

Transição	γ	μ_1	μ_2
Mapeamento do *dataset*	5	25,46	432,32
Redução do *dataset*	9	2,01	59,81

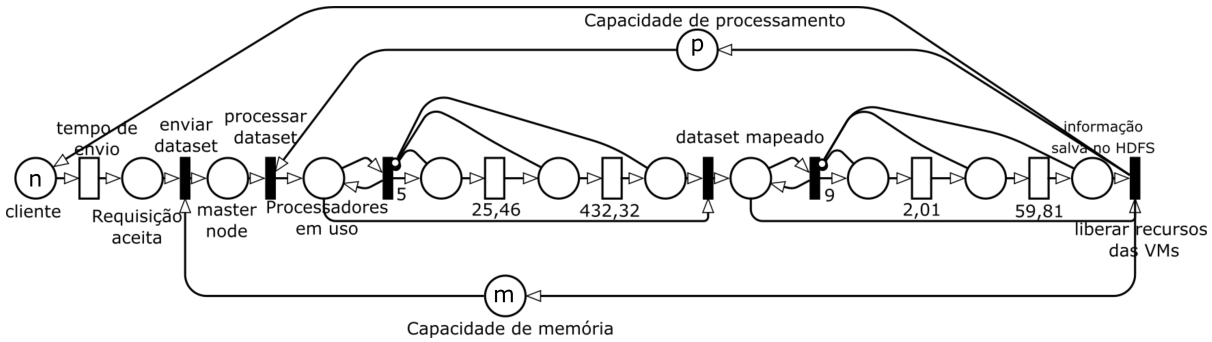

Figura 22 – Modelo de desempenho refinado (FONTE: Próprio autor)

Com o objetivo de definir a representação da carga de trabalho no modelo de desempnho, a Tabela 13 denota a representação da carga de trabalho no modelo de desempenho de acordo com a quantidade de marcações no lugar *Cliente*.

Tabela 13 – Representação da carga de trabalho no modelo de desempenho proposto

Quantidade de tokens	**Tamanho do *dataset* (*GB*)**
8	4
10	7
12	10

As capacidades de processamento e memória foram representadas no modelo de desempenho por meio de equações baseadas em regressão linear simples (GUPTA; GUTTMAN, 2014). A capacidade do *Hadoop* é definido pela oferta de serviço e pela quantidade de *data nodes* configurados em máquinas virtuais, em que as quantidades de *cores* de processamento e de memória são somadas no *cluster* (ALAPATI, 2016). No modelo

de desempenho, 12, 14 e 16 marcações no lugar *capacidade de processamento* denotam respectivamente 6, 12 e 24 *cores* de processamento, enquanto 14, 16 e 18 marcações no lugar *capacidade de memória* representam 12, 18 e 24 *GB* de memória disponível no *Hadoop cluster*.

Com isso, é possível ajustar estes pontos para uma reta que representa a quantidade de marcações nos lugares relativos à capacidade de processamento e de memória, para estimar o número de marcações para outras capacidades do *Hadoop cluster*, definido pelas ofertas de serviço e pelas quantidades de *data nodes* que somam suas capacidades no *cluster* (ALAPATI, 2016).

A regressão linear identificou a reta que melhor representa a quantidade de *cores* de processamento, com $R^2 = 96,4\%$, e a quantidade de *GB* de memória disponíveis, com $R^2 = 100\%$. Os valores de R^2 denotam que 96,4% da variação no número de marcações *tokens_cpu* é devido à variação do número de *cores* no modelo e que 100% de variação do número de marcações *tokens_mem* é devido à variação da quantidade de memória disponível no modelo de desempenho. Estes valores indicam que o número de marcações *tokens_cpu* e *tokens_mem* estão bem representadas na regressão linear simples. O parâmetro *cores* denota o número de *cores* disponíveis no *cluster* e *mem* representa o número de *GB* disponíveis no *cluster* instanciado na nuvem privada. Por exemplo, se forem considerados 8 cores e 24 *GB* de memória disponível, estes valores devem ser inseridos a partir da regressão linear simples para o mapeamento de *tokens*.

$$tokens\_cpu = 11 + 0,2143 \times cores \tag{6.1}$$

$$tokens\_mem = 10 + 0,3327 \times mem\ GB \tag{6.2}$$

Com base nestas equações, as marcações nos lugares *capacidade de processamento* e *capacidade de memória* no modelo de desempenho foram mapeadas da seguinte forma, como mostrado na Tabela 14. **M1** indica a quantidade de marcações no lugar ***capacidade de processamento*** e **M2** denota a quantidade de marcações no lugar ***capacidade de memória***.

Tabela 14 – Representação das capacidades de processamento e memória do *Hadoop cluster* no modelo de desempenho

Oferta de serviço	***Data nodes***	***Cores***	**Memória disponível(GB)**	**M1**	**M2**
Small	3	6	12	12	14
Small	5	10	20	13	16
Small	7	14	28	14	19
Medium	3	12	18	14	16
Medium	5	20	30	14	20
Medium	7	28	42	15	22
Large	3	24	24	16	18
Large	5	40	40	19	23
Large	7	56	56	20	25

6.2.12 Mapeamento de métricas de desempenho

As métricas de desempenho contempladas são as utilizações de processador e de memória dos *data nodes* configurados em máquinas virtuais da nuvem privada. As métricas de utilização de processador e de memória foram calculads pelas Equações apresentadas na Tabela 3 do Capítulo 5.

6.2.13 Validação e Verificação do modelo de desempenho

A validação do modelo de desempenho proposto consiste na comparação entre os valores das métricas de utilização de processador e de memória medidas do ambiente *Big Data* e dos valores das métricas de utilização de processador e de memória calculados através das Equações apresentadas na Tabela 3 do Capítulo 5. Nesse trabalho, essa comparação foi realizada através de testes T emparelhado (GUPTA; GUTTMAN, 2014).

A Tabela 7 denota as configurações dos cenários adotados para a validação do modelo de desempenho, indicando qual oferta de serviço, quantidade de *data nodes* e tamanho de *dataset* é avaliado e a Figura 23 ilustra como ocorreu o processo de validação dos 27 cenários(Tabela 7) gerados pelo planejamento de experimentos.

Um subproblema inicial é validado (ambientes com 3 *data nodes*)(Cenários 1 até 9), e estende-se a verificação do modelo de desempenho para outros subproblemas no ambiente

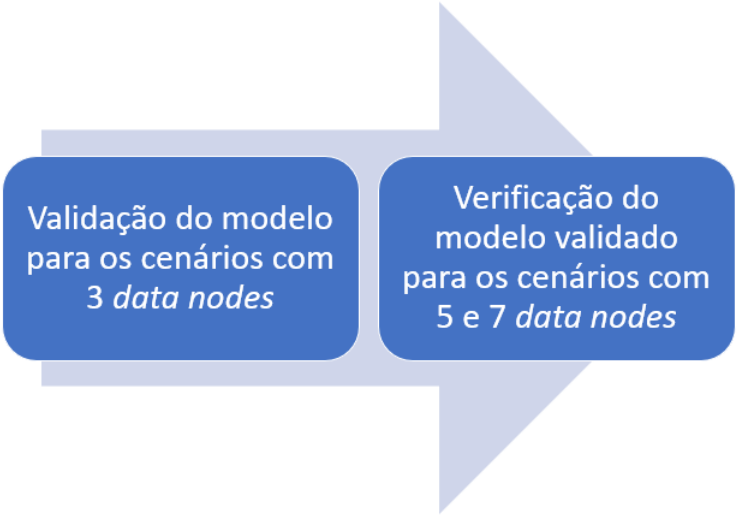

Figura 23 – Processo adotado para a validação e verificação dos 27 cenários (FONTE: Próprio autor)

sob avaliação (ambientes com 5 e 7 *data nodes* (JAIN, 1991))(Cenários 10 até 27).

Este processo pode ser utilizado para ambientes com maiores quantidades de *data nodes*, como 9, 10, 11, ..., *n data nodes*, conforme a capacidade da nuvem privada em instanciar mais máquinas virtuais para serem configuradas no *Hadoop cluster*. Este aspecto é definido pelo fator *Número de data nodes* adotado no planejamento de experimentos.

A Tabela 15 apresenta as métricas de utilização de processador e de memória medidas nos *data nodes* configurados em máquinas virtuais instanciadas na nuvem privada e as métricas utilização de processador e utilização de memória calculadas através do modelo de desempenho proposto para os Cenários 1 até 9(Ver Tabela 7).

A Tabela 16 apresenta os resultados dos testes T emparelhado considerando as métricas de utilização de processador e de memória dos *data nodes* instanciados na nuvem privada para a análise do conjunto de dados (GUPTA; GUTTMAN, 2014).

Pode-se observar que o intervalo de confiança considerando ambas as métricas incluíram o zero. Isto sugere que não há evidências estatísticas de que existem diferenças significativas entre os resultados obtidos nas medições e obtidos através das métricas mapeadas no modelo de desempenho. Outro aspecto importante são os Valores-P dos dois testes, que foram acima de 0,05, referente à confiança de 95%, indicando a forte semelhança entre os resultados comparados (GUPTA; GUTTMAN, 2014). Os testes T emparelhados foram realizados através do *software* estatístico Minitab (MINITAB, 2018) com o objetivo de avaliar se o modelo de desempenho proposto representa os diferentes cenários contemplados nas medições. Os Cenários com 5 e 7 *data nodes* são avaliados para

Tabela 15 – Métricas Utilização de Processador e Utilização de memória medidas e calculadas através do modelo de desempenho

Cenários	**%Uproc-medição**	**%Uproc-modelo**	**%Umem-medição**	**%Umem-modelo**
1	77,03	74,99	65,11	49,99
2	80,55	66,65	75,40	57,13
3	82,36	83,32	80,08	71,42
4	69,67	71,41	48,22	62,49
5	71,73	57,13	60,11	49,99
6	72,47	92,84	78,17	81,24
7	58,64	62,49	39,86	55,55
8	60,84	74,99	52,12	66,65
9	62,12	68,74	40,34	66,55

a verificação do modelo de desempenho.

Tabela 16 – Resultados dos testes T emparelhado das métricas de desempenho para os cenários com 3 *data nodes*

Métrica	**Intervalo de confiança(95%)**	**P-valor**
Utilização de processador(%)	(-10,72; 6,95)	0,636
Utilização de memória(%)	(-14,72; 9,89)	0,663

A Tabela 17 apresenta as utilizações de processador e de memória provenientes das medições e do modelo de desempenho, com o intuito de verificar o modelo de desempenho proposto para os cenários com 5 *data nodes* e 7 *data nodes* configurados na infraestrutura de nuvem privada para a análise do *data set*.

A Tabela 18 denota os resultados dos testes T emparelhados para os cenários contemplando 5 e 7 *data nodes* no *cluster* para a análise dos conjuntos de dados. Pode-se observar que todos os intervalos de confiança incluem o zero e que ainda todos os P-valores são maiores que 0,05. Portanto, o modelo de desempenho proposto neste trabalho foi verificado para os Cenários com 5 e 7 *data nodes*(Cenários 10 até 27 da Tabela 7) definidos no planejamento de experimentos.

Tabela 17 – Utilização de processador e memória dos *data nodes* configurados na nuvem privada para os cenários com 5 e 7 *data nodes*

Cenário	**%Uproc-medição**	**%Uproc-modelo**	**%Umem-medição**	**%Umem-modelo**
10	75,30	76,91	56,23	62,49
11	79,01	61,53	67,07	49,99
12	82,74	76,91	79,61	62,49
13	78,45	71,42	40,40	49,99
14	76,91	85,70	49,07	59,99
15	75,86	78,56	61,57	54,99
16	62,82	52,62	37,46	43,47
17	60,91	63,15	42,88	52,16
18	61,37	63,15	55,63	52,16
19	80,27	71,42	47,85	52,62
20	81,25	80,44	31,27	42,00
21	82,41	85,70	63,28	63,15
22	70,17	62,49	35,64	21,73
23	70,86	54,55	53,32	36,36
24	77,36	62,49	58,46	43,47
25	55,29	49,99	33,94	39,99
26	59,19	59,99	34,99	47,99
27	59,45	64,99	40,34	51,99

Tabela 18 – Resultados dos testes T emparelhado das métricas de desempenho para os cenários com 5 e 7 *data nodes*

Métrica	**Intervalo de confiança(95%)**	**P-valor**
Utilização de processador(%)	(-0,11; 7,63)	0,056
Utilização de memória(%)	(-5,54; 5,76)	0,968

Portanto, o modelo está validado para os Cenários 1 até 9 da Tabela 7, que representam os ambientes com 3 *data nodes* e verificado para os ambientes com 5 e 7 *data nodes*(Cenários 10 até 27 da Tabela 7). O modelo pode ser adotado para avaliar diferentes cenários como considerar maiores cargas de trabalho e/ou maiores capacidades de ofertas de serviço e *data nodes*.

6.2.14 Planejamento de experimentos para a aplicação *Big Data* na nuvem privada

A medição contemplou cenários que processaram *datasets* de até 10 *GB* e o modelo de desempenho validado e verificado avaliou ambientes considerando maiores cargas de trabalho, que são representadas pelo fator Tamanho do *dataset* definido no planejamento de experimentos.

A Tabela 19 apresenta os fatores e níveis estabelecidos para o planejamento de experimentos de novos cenários. Os níveis dos fatores Oferta de serviço e Número de *data nodes* permenecer os mesmos dos adotados no planejamento de experimentos do ambiente de medição enquanto que os níveis 15, 20, 25 e 30 GB foram considerados para o fator Tamanho do *dataset*.

Tabela 19 – Fatores e níveis adotados para avaliar cenários com maiores tamanhos de *dataset*

Fatores	**Níveis (1, 2, 3, 4)**			
Oferta de serviço	Small	Medium	Large	-
Número de data nodes	3	5	7	-
Tamanho do dataset(GB)	15	20	25	30

6.2.15 Análise de Novos Cenários

O objetivo deste estudo é avaliar a carga de trabalho máxima suportada pelo *Hadoop cluster* com 3, 5 e 7 *data nodes* configurados na infraestrutura de nuvem privada, conforme apresentando na Tabela 19. Além disso, este estudo de caso também tem o objetivo de avaliar os custos de infraestrutura, de software e de energia elétrica desta infraestrutura de nuvem privada considerando estas cargas de trabalho.

A utilização de processador($Uprocessador$) e a utilização de memória($Umem$), são calculados considerando todos os cenários gerados(todas as ofertas de serviço e número de *data nodes*) (MINITAB, 2018)(Ver Tabela 19). $Uprocessador$ e $Umem$ foram calculados no modelo de desempenho validado e verificado, no qual maiores cargas são aplicadas(15 *GB*, 20 *GB*, 25 *GB* e 30 *GB*).

As Tabelas 20, 21 e 22 apresentam a utilização média de processadores e de memória de *data nodes* configurados em máquinas virtuais da nuvem privada quando maiores intensidades de carga de trabalho são aplicadas, considerando ambientes *Big Data* com 3, 5 e 7 *data nodes*, respectivamente. Para ambientes com 3 *data nodes*, pode-se notar que todas as ofertas de serviço apresentaram uma saturação do recurso de processamento e de memória, uma vez que $Uprocessador$ e $Umem$ apresentaram valores maiores que 90% em quase todos os cenários (CILIENDO, 2005), conforme mostrado na Tabela 20. Isto sugere que ambientes com um número menor ou igual que 3 *data nodes* não são recomendados para *datasets* de tamanho acima de 10 *GB*.

Tabela 20 – Utilizações de processador e de memória dos *data nodes* configurados na nuvem para os Cenários com 3 *data nodes*

Cenário	**Carga de trabalho(GB)**	**Oferta de serviço**	$Uprocessador$(%)	$Umem$(%)
1	15	Small	99,999	92,849
2		Medium	99,999	91,743
3		Large	99,999	88,883
4	20	Small	99,999	99,999
5		Medium	99,999	99,993
6		Large	99,999	94,438
7	25	Small	99,99999	99,999
8		Medium	99,999	99,998
9		Large	99,999	99,994
10	30	Small	99,999	99,999
11		Medium	99,999	99,998
12		Large	99,999	99,999

Os Cenários avaliados no modelo de desempenho considerando 5 *data nodes*, apresentaram uma tendência de saturação para o recurso de processamento (CILIENDO, 2005) quando as menores ofertas de serviço (*small* e *medium*) são adotadas para configuração dos *data nodes*.

A memória dos *data nodes* configurados em máquinas virtuais da nuvem privada apresentaram uma tendência para saturar para a oferta de serviço *small* e para *data sets* de tamanho acima de 25 *GB* (CILIENDO, 2005). Para ambientes *Big Data* do *Hadoop*

com 5 *data nodes*, é recomendado adotar ofertas de serviço mais robustas para evitar a saturação de processadores e de memórias.

A Tabela 22 denota que para cargas de trabalho de 30 *GB* o recurso de processamento satura para todas as ofertas de serviço e que o recurso de memória consegue suportar cargas de 15 *GB*, mas que pode outro lado, apresentam uma maior tendência a saturar quando menores ofertas de serviço são consideradas para os *data nodes*.

Tabela 21 – Utilizações de processador e memória dos *data nodes* configurados na nuvem para os Cenários com 5 *data nodes*

Cenário	**Carga de trabalho(GB)**	**Oferta de serviço**	$\%U_{processador}$	$\%U_{mem}$
13	15	Small	99,992	81,243
14		Medium	99,999	74,994
15		Large	68,415	56,517
16	20	Small	99,999	99,993
17		Medium	99,999	79,994
18		Large	84,205	69,560
19	25	Small	99,999	99,999
20		Medium	99,999	89,994
21		Large	94,731	78,256
22	30	Small	99,999	99,999
23		Medium	99,999	99,999
24		Large	99,999	99,9997

As Tabelas 20 e 21 e 22 mostram que o aumento da oferta de serviço aliado ao aumento da quantidade de *data nodes* no *Hadoop cluster*, proporciona uma redução na utilização de processadores e memória quando maiores cargas de trabalho são aplicadas ao ambiente *Big Data* na nuvem privada.

Visto que a infraestrutura e o conjunto de *software* permanecem os mesmos dos cenários considerados nas medições, os seus custos permanecem com o valor de R$ 16.773,77 e R$ 0,00 respectivamente. O custo de consumo de energia de acordo com a Equação 5.3 depende do tempo de execução, consumo de energia elétrica e do valor kWh(uma constante).

Tabela 22 – Utilizações de processador e memória dos *data nodes* configurados na nuvem para os Cenários com 7 *data nodes*

Cenário	**Carga de trabalho(GB)**	**Oferta de serviço**	$\%U_{processador}$	$\%U_{mem}$
25	15	Small	92,949	68,415
26		Medium	91,9931	68,995
27		Large	79,9948	59,995
28	20	Small	99,998	84,205
29		Medium	99,999	79,994
30		Large	79,999	63,995
31	25	Small	100,000	94,731
32		Medium	99,998	89,994
33		Large	99,999	71,995
34	30	Small	100,000	99,998
35		Medium	99,999	99,999
36		Large	95,972	95,995

Logo, para avaliar o custo associado às cargas de trabalho de 15, 20, 25 e 30 *GB*, é preciso que os tempos de execução e o consumo de energia nestes cenários sejam estimados, cosiderando as diferentes ofertas de serviço provisionadas pela nuvem privada.

Para identificar a equação que melhor representa os tempos de execução, em segundos, considerando os fatores Oferta de serviço, Número de *data nodes* e Tamanho do *data set*, a análise de regressão foi realizada através do Minitab (MINITAB, 2018) *software* estatístico de acordo com os cenários gerados pelo planejamento de experimentos adotados nas medições(Tabela 7) em que a oferta de serviço é composta pelo número de *cores* de processamento e quantidade de memória para cada oferta de serviço(Tabela 19). A Equação 6.3 denota a equação de regressão com o parâmetro $R^2 = 86,45\%$, indicando uma boa associação entre a variável de resposta(Tempo de execução da aplicação *Big Data*) e os fatores. Este parâmetro denota que 86,45% da variação dos tempos de execução é devido à variação destes fatores considerados.

$$\Delta T = 379\ s - 25,6\ s \times NDN + 67,2\ s \times TD + 23,5\ s \times cores - 60,3\ s \times mem \quad (6.3)$$

Em que NDN=Número de *data nodes* do *Hadoop cluster*; TD=Tamanho do *data set*; *cores*=quantidade de *cores* de processamento em cada oferta de serviço; *mem*=quantidade

de memória definido para cada oferta de serviço.

Para avaliar a análise de regressão para o consumo de energia elétrica do ambiente *Big Data* na nuvem privada(em *kiloWatts*), o procedimento similar para a métrica Tempo de execução foi realizado. A Equação 6.4 apresenta a equação de regressão calculado pelo Minitab (MINITAB, 2018) com um parâmetro de $R^2 = 67,88\%$.

$$kW = -3,8kW + 6,4kW \times NDN + 0,2kW \times TD - 1,8kW \times cores + 7,7kW \times mem \tag{6.4}$$

Com base nas Equações 6.3 e 6.4, a Tabela 23 apresenta os custos considerando cargas de trabalho de 15, 20, 25 e 30 *GB*.

Tabela 23 – Custo de energia total estimado para cargas de trabalho de 15, 20, 25 e 30 *GB* em cenários de 7 *data nodes* configurados em máquinas virtuais da nuvem privada

Cenário	**Carga de trabalho(GB)**	**Oferta de serviço**	**Energia(R$)**	**Custo total(R$)**
25	15	Small	0,02	16.773,79
26		Medium	0,03	16.773,80
27		Large	0,03	16.773,80
28	20	Small	0,02	16.773,79
29		Medium	0,03	16.773,80
30		Large	0,04	16.773,81
31	25	Small	0,03	16.773,80
32		Medium	0,04	16.773,81
33		Large	0,04	16.773,81
34	30	Small	0,03	16.773,80
35		Medium	0,04	16.773,81
36		Large	0,05	16.773,82

O custo total(Custo total(R$)) e o custo de consumo de energia(Energia(R$)) são calculados pelas Equações 5.1 e 5.3, respectivamente. Os cenários com 7 *data nodes* são considerados na tabela, pois, o objetivo é avaliar ambientes mais robustos, que nesse caso são os cenários com o maior número de *data nodes* que a infraestrutura de nuvem suportou. A Tabela 23 denota o custo total do ambiente *Big Data* implantado na infraestrutura de nuvem privada. Pode-se notar que o custo total é impactado principalmente pelo custo

de infraestrutura da nuvem privada para tempos de execução pequenos. Para cenários que analisam *data sets* que demandam longos períodos de tempo para serem processados, o custo de energia elétrica do ambiente *Big Data* pode apresentar um maior impacto nos custos totais. Pode-se tomar como exemplo, o caso em que *Petabytes* de dados são processados demandando um longo período de tempo para serem completamente processados na nuvem privada.

6.3 Considerações Finais

O presente capítulo demonstrou a aplicabilidade da metodologia proposta para avaliação de desempenho e de custo de aplicações *Big Data* em ambientes de nuvem privada em um cenário real. As medições e os modelos de desempenho e de custo identificaram fatores impactantes nestes ambientes, no qual cargas de trabalho de 15, 20, 25, e 30 *GB* foram avaliadas quanto á utilização de processadores, utilização de memória de *data nodes* do *Hadoop* e o custo, considerando as ofertas de serviço da nuvem privada.

7 Conclusão

A geração de dados cada vez mais volumosos e de diferentes tipos necessitam de tecnologias que processem estes dados de forma eficiente, sem *overheads* e gargalos, e de forma eficaz, analisar dados do *data set* e gerar a informação que deseja a partir do *datasets*. A computação em nuvem surge neste contexto oferecendo um *pool* de recursos para o processamento de diferentes *data sets* através de mecanismos distribuídos, como *Hadoop* implantado em uma nuvem privada, com *clusters* processandos os dados. Avaliar como a nuvem privada se comporta ao executar cargas de trabalho é um desafio para a comunidade científica e diferentes metodologias, modelos de desempenho e modelos de custos estão sendo propostos.

Este trabalho propôs uma estratégia baseada em uma metodologia e modelos para avaliação de desempenho e de custo de aplicações *Big Data* em ambientes de nuvem privada. A metodologia proposta que contempla atividades relacionadas a geração de um modelo de desempenho, de modelos de custo e o planejamento de experimentos considerando diferentes fatores que impactaram no desempenho e no custo destas aplicações. O planejamento de experimentos considera fatores como ofertas de serviço da nuvem privada, número de *data nodes* do *Hadoop cluster* e tamanho do conjunto de dados a ser analisado através de mecanismos distribuídos. As variáveis de resposta adotadas neste trabalho para avaliar o impacto dos fatores são os tempos de execução, utilização de processador, utilização de memória e consumo de energia elétrica.

Um estudo de caso demonstrou a aplicabilidade da metodologia proposta em um cenário real de nuvem privada, no qual *Hadoop clusters* foram implantados em uma nuvem privada, com diferentes ofertas de serviço, números de *data nodes* e cargas de trabalho. O *data set* foi gerado a partir de opiniões de usuários da rede social *Twitter* sobre as eleições presidenciais de 2018, com a identificação de termos(#*hashtags* mais citados relacionados aos candidatos.

Os resultados das medições aliadas à análise de sensibilidade demonstraram que para os tempos de execução da aplicação *Big Data* na nuvem privada, o tamanho do *data set* é o fator mais impactantes nesta métrica de desempenho. Os tempos de execução apresentam um impacto maior relacionado ao tamanho do *data set* processado comparado à oferta de serviço da nuvem privada e a quantiade de *data nodes* configurados em máquinas

virtuais. Em relação à utilização de processador dos *data nodes* configurados em máquinas virtuais da nuvem privada, a oferta de serviço apresentou o maior impacto comparado aos outros fatores avaliados. Isto mostra que aumentar a capacidade da nuvem privada é uma alternativa recomendada, visto que ofertas de serviço mais robustas podem ser implantadas para as máquinas virtuais. A utilização de memória das máquinas virtuais apresentou uma equidade entre os três fatores avaliados. Por fim, a análise do consumo de energia nos experimentos, demonstrou que a oferta de serviço e a quantidade de *data nodes* configurados são os que mais impactam nesse fator. O tamaho do *dataset* quase não apresentou impacto. Pode-se afirmar que o consumo de energia em aplicações *Big Data* pode ser decrementado a medida que a capacidade da nuvem privada e ofertas de serviço mais robustas são adotadas para estes *data nodes*. O custo total foi composto de custos relacionados à infraestrutura de nuvem privada, custos associados ao consumo de energia elétrica da aplicação *Big Data* e custos de *software*. O custo de infraestrutura de nuvem privada apresentou um valor de R$ 16.773,77, enquanto que os custos de consumo de energia elétrica e de *software* foram bem menores, indicando que o custo de investimento na nuvem privada tem um valor considerável para implantacação do ambientes para o processamento de dados.

Com o modelo de desempenho e os modelos de custo propostos neste trabalho, analistas de ambientes *Big Data*, analistas de desempenho, analistas de infraestrutura dentre outros, podem avaliar diversos cenários de *Data Analytics* e *Data Analysis* que utilizam computação em nuvem para prover os recursos computacionais necessários para a análise dos dados.

O modelo de desempenho é baseado em redes de Petri estocásticas e prorporciona a avaliação de utilização de processadores e de memória dos *data nodes* na nuvem privada. O mecanismo *MapReduce* foi modelado considerando os tempos que a aplicação demanda para mapear os dados e os tempos referentes à geração das informações resultantes da análise dos *data sets*. O refinamento do modelo de desempenho consiste na identificação e caracterização destes tempos(mapeamento e redução) e na validação do modelo de desempenho.

A análise dos modelos de custo propostos demonstraram a viabilidade dos mesmos no custo de infraestrutura de nuvem privada e que é o custo mais impactante considerando este ambiente. Quando os tempos de execução das cargas de trabalho geradas pela análise

de *data sets* não são muito extensos, o custo associado ao consumo de energia elétrica é pequeno para cada aplicação *Big Data* na nuvem privada.

Na análise de novos cenários, os modelos de desempenho e de custo propostos foram adotados considerando maiores tamanhos de *data sets* do que os considerados para os experimentos. O modelo de desempenho validado foi adaptado para maiores cargas de trabalho e demonstrou que para cenários com apenas 3 *data nodes* configurados na nuvem privada, nenhuma oferta de serviço suportou cargas maiores do que 15 *GB*. Os cenários com 5 e 7 *data nodes* apresentaram *overheads* para as menores ofertas de serviço da nuvem privada. Ainda baseado no modelo de desempenho, foi concluído que a infraestrutura de nuvem privada, não suporta a análise de *data sets* a partir de 30 *GB*. O modelo de desempenho também concluiu que *Hadoop clusters* com menor quantidade de *data nodes* configurados na nuvem privada, possuem maiores chances de apresentar *overhead* dos recursos de memória destes *data nodes*. Para ambientes com maiores quantidades de *data nodes*, o modelo de desempenho demonstrou que a probabilidade de processadores e memória dos *data nodes* configurados apresentarem *overhead* nas máquinas virtuais diminui cerca de 15%, mesmo que o tamanho do *data set* seja incrementado. Além disso, a utilização de memória do *Hadoop cluster* para tamanhos de *datasets* até 25 *GB*, podem ser processados com as atuais ofertas de serviço e quantidades de *data nodes* do ambiente real, sem que haja um aumento na taxa de utilização memória.

7.1 Contribuições

Este trabalho tem como contribuição principal uma estratégia para a avaliação de desempenho e de custo de aplicações *Big data* em infraestruturas de nuvens privadas. Um modelo de desempenho, que representa *data sets* sendo processados pelo *Hadoop*, é proposto para avaliar a utilização de recursos da aplicação e modelos de custos são propostos para avaliar custos de infraestrutura de nuvem privada, consumo de energia elétrica e de aquisição de *software*. As demais contribuições são descritas a seguir:

- **Metodologia Para a Avaliação de Aplicações *Big Data* em Infrestruturas de Nuvens Privadas** - o objetivo da metodologia é propor uma estratégia para avaliação de desempenho e custo de ambientes *Big Data* suportados por

infraestruturas de nuvem privada. A metodologia proposta pode ser adotada em diversos ambientes *Big Data* que possuem uma infraestrutura de nuvem privada para provisionar os recursos computacionais considerando a análise de *data sets*.

- **Planejamento de Experimentos** - o planejamento de experimentos fatorial demonstrou os fatores mais impactantes para métricas de desempenho e de custo como tempos de execução, utilizações de processadores e de memória, e consumo de energia demandado pelo ambiente *Big Data* e custos da nuvem privada. Isto contribuiu para o gerenciamento do ambiente *Big Data* na nuvem privada, uma vez que os fatores que impactam no desempenho e custo deste ambiente podem ser identificados e controlados.
- **Modelo de Desempenho** - um modelo baseado em redes de Petri estocásticas foi concebido para avaliar e calcular a utilização de processador e de memória dos *data nodes* instanciados na nuvem privada para ambientes *Big Data* de diversos cenários, considerando diferentes ofertas de serviço da nuvem privada e quantidades de *data nodes* configurados em máquinas virtuais assim como diferentes tamanhos do conjunto de dados. Com o modelo de desempenho, pode-se prever métricas, como utilização de processadores e de memória.
- **Modelos de Custo** - modelos de custos associados a ambientes *Big Data* em nuvens privadas foram propostos considerando o custo de infraestrutura de nuvem privada, custo de consumo de energia elétrica e custo de *software* relativos à estes ambientes.

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;
- A avaliação contempla apenas a nuvem privada (MELL et al., 2011; ERL et al., 2013);
- Apenas o ambiente *Hadoop* foi considerado para a avaliação de desempenho e de custo de aplicações *Big Data* em ambientes de nuvem privada.
- A platforma Cloudstack (CLOUDSTACK, 2018) foi adotada para gerenciamento e

orquestração da nuvem privada devido à sua flexibilidade quanto ao gerenciamento de ofertas de serviço, alocação de recursos e rapidez na instanciação de máquinas virtuais.

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

7.3 Trabalhos Futuros

Os seguintes trabalhos futuros podem ser estendidos a partir desta dissertação:

- Outros ambientes *Big Data* podem ser adotados para a avaliação de desempenho e custo;
- Pode-se implementar um *benchmark* e automatizar a avaliação de desemepenho e custo de aplicações *Big Data* em ambientes de nuvem privada. Com a metodologia e os modelos propostos neste trabalho, cenários com capacidades mais robustas e maiores cargas de trabalho podem ser avaliados ao analisar diferentes conjuntos de dados;
- Avaliar a utilização de recursos quando diversos tipos de *datasets* são processados pelo *Hadoop*, sejam dados de redes sociais, dados relativos ao gerenciamento de saúde(*healthcare data*) (HASSAN et al., 2019), dados de sensores capturados por dispositivos *IoT* (SOLANKI et al., 2019), dentre outros diversos tipos de dados gerados por diferentes fontes (ASSUNÇÃO et al., 2015; HASHEM et al., 2015; ALAPATI, 2016; OUSSOUS et al., 2018);
- Nuvens públicas e híbridas podem ser adotadas para provisionar os recursos necessários e avaliar o seu desempenho e custo comparado à nuvem privada considerada neste trabalho;
- Avaliar a confiabilidade e disponibilidade de *Hadoop clusters* implantados em ambientes de nuvem considerando gerenciamento de falhas.
- Avaliar a performabilidade de ambientes *Big Data* em nuvens privadas e estimar degradação de desempenho devido às falhas de um k número de *data nodes* no *cluster* e prever definições de *SLAs*(*service level agreements*) entre provedores de serviços de computação em nuvem e seus clientes.

Referências

ADHIKARI, B. K.; ZUO, W.; MAHARJAN, R. A performance analysis of openstack cloud vs real system on hadoop clusters. In: IEEE. **2017 IEEE International Conference on Edge Computing (EDGE)**. [S.l.], 2017. p. 194–201.

ALAPATI, S. R. **Expert Hadoop Administration: Managing, Tuning, and Securing Spark, YARN, and HDFS**. [S.l.]: Addison-Wesley Professional, 2016.

APACHE. **Apache Hadoop**. [S.l.], 2018. Disponível em: <https://hadoop.apache.org/>. Acesso em: 27 mar. 2018.

ARDUINO. **Arduino Create simplifies building a project as a whole, without having to switch between different tools to manage all the aspects of whatever you're making.** [S.l.], 2018. Disponível em: <https://create.arduino.cc/>. Acesso em: 02 nov. 2018.

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

BARTON, R. R. **Graphical methods for the design of experiments**. [S.l.]: Springer Science & Business Media, 2012. v. 143.

CELPE. **PREÇOS FINAIS DE ENERGIA ELÉTRICA - GRUPO A**. [S.l.], 2018. Disponível em: <http://servicos.celpe.com.br/residencial-rural/Documents/tarifas/CELPE_TARIFAS_E_PRECOS_FINAIS_GRUPO_A_MARCO_2019.pdf>. Acesso em: 10 nov. 2018.

CENTOS. **The CentOS Project**. [S.l.], 2018. Disponível em: <https://www.centos.org/>. Acesso em: 22 mar. 2017.

CHEN, S.; RODERO, I. Understanding behavior trends of big data frameworks in ongoing software-defined cyber-infrastructure. In: ACM. **Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies**. [S.l.], 2017. p. 199–208.

CHINOSI, M.; TROMBETTA, A. Bpmn: An introduction to the standard. **Computer Standards & Interfaces**, Elsevier, v. 34, n. 1, p. 124–134, 2012.

CILIENDO, E. Tuning red hat enterprise linux on ibm eserver xseries servers. **IBM Redpaper**, 2005.

CLOUDSTACK. **Apache CloudStack™ Open Source Cloud Computing™**. [S.l.], 2018. Disponível em: <https://cloudstack.apache.org/>. Acesso em: 03 dez. 2017.

COLUMBUS. **10 Charts That Will Change Your Perspective Of Big Data's Growth**. [S.l.], 2018. Disponível em: <https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#3999643f2926>. Acesso em: 25 nov. 2017.

DATANYZE. **TOP COMPETITORS OF APACHE SPARK IN DATANYZE UNIVERSE**. [S.l.], 2018. Disponível em: <https://www.datanyze.com/market-share/big-data-processing/apache-spark-market-share>. Acesso em: 18 ago. 2017.

DEAN, J.; GHEMAWAT, S. Mapreduce: a flexible data processing tool. **Communications of the ACM**, ACM, v. 53, n. 1, p. 72–77, 2010.

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

ERL, T.; PUTTINI, R.; MAHMOOD, Z. **Cloud computing: concepts, technology & architecture**. [S.l.]: Pearson Education, 2013.

EUCALYPTUS. **Official Documentation for Eucalyptus Cloud**. [S.l.], 2017. Disponível em: <https://docs.eucalyptus.cloud/eucalyptus/4.4.5/index.html>. Acesso em: 23 nov. 2017.

FEITELSON, D. G. **Workload modeling for computer systems performance evaluation**. [S.l.]: Cambridge University Press, 2015.

FEMMINELLA, M.; PERGOLESI, M.; REALI, G. Performance evaluation of edge cloud computing system for big data applications. In: IEEE. **2016 5th IEEE International Conference on Cloud Networking (Cloudnet)**. [S.l.], 2016. p. 170–175.

GATES, A.; DAI, D. **Programming pig: Dataflow scripting with hadoop**. [S.l.]: "O'Reilly Media, Inc.", 2016.

GUPTA, B. C.; GUTTMAN, I. **Statistics and probability with applications for engineers and scientists**. [S.l.]: John Wiley & Sons, 2014.

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

HASSAN, M. K.; DESOUKY, A. I. E.; ELGHAMRAWY, S. M.; SARHAN, A. M. Big data challenges and opportunities in healthcare informatics and smart hospitals. In: **Security in Smart Cities: Models, Applications, and Challenges**. [S.l.]: Springer, 2019. p. 3–26.

HIPS. **HiPS : Hierarchical Petri net Simulator**. [S.l.], 2018. Disponível em: <http://hips-tools.sourceforge.net/wiki/>. Acesso em: 09 set. 2018.

IBM. **IBM Big Data Analytics: *insights* sem limites**. [S.l.], 2018. Disponível em: <https://www.ibm.com/br-pt/it-infrastructure/solutions/big-data>. Acesso em: 24 jul. 2017.

INA. **INA Integrated Net Analyzer**. [S.l.], 2018. Disponível em: <https://www2.informatik.hu-berlin.de/~starke/ina.html>. Acesso em: 25 nov. 2018.

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

JEFF, G. **Package 'twitteR'**. [S.l.], 2016. Disponível em: <https://cran.r-project.org/web/packages/twitteR/twitteR.pdf>. Acesso em: 10 nov. 2017.

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: IEEE. **2015 IEEE International Conference on Cluster Computing**. [S.l.], 2015. p. 650–657.

KAMTEKAR, K. **Performance Modeling of Big Data**. [S.l.]: May, 2015.

KIM, Y.-H.; HUH, E.-N. Towards the design of a system and a workflow model for medical big data processing in the hybrid cloud. In: IEEE. **2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)**. [S.l.], 2017. p. 1288–1291.

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

LONGEN, A. **CentOS vs Ubuntu: saiba qual sistema escolher para seu servidor web**. [S.l.], 2018. Disponível em: <https://www.hostinger.com.br/tutoriais/centos-vs-ubuntu-qual-escolher-para-servidor-web/>. Acesso em: 07 mai. 2017.

LOVAS, R.; NAGY, E.; KOVÁCS, J. Cloud agnostic big data platform focusing on scalability and cost-efficiency. **Advances in Engineering Software**, Elsevier, v. 125, p. 167–177, 2018.

LU, Z.; WANG, X.; WU, J.; HUNG, P. C. Instechah: Cost-effectively autoscaling smart computing hadoop cluster in private cloud. **Journal of Systems Architecture**, Elsevier, v. 80, p. 1–16, 2017.

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

MACIEL, P.; MATOS, R.; SILVA, B.; FIGUEIREDO, J.; OLIVEIRA, D.; FÉ, I.; MACIEL, R.; DANTAS, J. Mercury: performance and dependability evaluation of systems with exponential, expolynomial, and general distributions. In: IEEE. **2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC)**. [S.l.], 2017. p. 50–57.

MAHESHWARI, A. K. Application of big data to smart cities for a sustainable future. **Handbook of Engaged Sustainability**, Springer, p. 1–24, 2018.

MALIK, M.; NAAZ, S.; ANSARI, I. R. Sentiment analysis of twitter data using big data tools and hadoop ecosystem. In: SPRINGER. **International Conference on ISMAC in Computational Vision and Bio-Engineering**. [S.l.], 2018. p. 857–863.

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

MARSAN, M. A.; BALBO, G. **Modelling with generalized stochastic Petri nets**. [S.l.]: John Wiley & Sons, Inc., 1994.

MASSOBRIO, R.; NESMACHNOW, S.; TCHERNYKH, A.; AVETISYAN, A.; RADCHENKO, G. Towards a cloud computing paradigm for big data analysis in smart cities. **Programming and Computer Software**, Springer, v. 44, n. 3, p. 181–189, 2018.

MELL, P.; GRANCE, T. et al. The nist definition of cloud computing. Computer Security Division, Information Technology Laboratory, National . . . , 2011.

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

MINITAB. **Software estatístico poderoso que todos podem usar**. [S.l.], 2018. Disponível em: <http://www.minitab.com/pt-br/products/minitab/>. Acesso em: 12 set. 2018.

MONTGOMERY, D. C. **Design and analysis of experiments**. [S.l.]: John wiley & sons, 2017.

MURATA, T. Petri nets: Properties, analysis and applications. **Proceedings of the IEEE**, IEEE, v. 77, n. 4, p. 541–580, 1989.

NIST. **NIST Cloud Computing Program - NCCP**. [S.l.], 2018. Disponível em: <https://www.nist.gov/programs-projects/nist-cloud-computing-program-nccp>. Acesso em: 25 nov. 2017.

OPENSTACK. **What is OpenStack?** [S.l.], 2017. Disponível em: <https://www.openstack.org/software/>. Acesso em: 23 nov. 2017.

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

PALIAN, J. **How the Cost of Cloud Computing is Calculated**. [S.l.], 2018. Disponível em: <https://www.expedient.com/blog/how-the-cost-of-cloud-computing-is-calculated/>. Acesso em: 13 nov. 2018.

R. **The R Project for Statistical Computing**. [S.l.], 2018. Disponível em: <https://www.r-project.org/>. Acesso em: 08 jul. 2018.

RUIZ, M. C.; CALLEJA, J.; CAZORLA, D. Petri nets formalization of map/reduce paradigm to optimise the performance-cost tradeoff. In: IEEE. **2015 IEEE Trustcom/BigDataSE/ISPA**. [S.l.], 2015. v. 3, p. 92–99.

RUIZ, M. C.; CAZORLA, D.; PÉREZ, D.; CONEJERO, J. Formal performance evaluation of the map/reduce framework within cloud computing. **The Journal of Supercomputing**, Springer, v. 72, n. 8, p. 3136–3155, 2016.

SILVA, J. d. S. **Uma metodologia para avaliação do consumo de energia de aplicações baseadas em ambientes de computação móvel em nuvem**. Dissertação (Mestrado), 2017. Departamento de Estatística e Informática. Disponível em: <http://www.tede2.ufrpe.br:8080/tede2/handle/tede2/7866>.

SOLANKI, V. K.; DÍAZ, V. G.; DAVIM, J. P. **Handbook of IoT and Big Data**. [S.l.]: CRC Press, 2019.

STATISTA. **Size of Hadoop and Big Data markets worldwide in 2015 and 2020**. [S.l.], 2018. Disponível em: <https://www.statista.com/statistics/587051/worldwide-hadoop-bigdata-market/>. Acesso em: 26 jun. 2017.

TANENBAUM, A. S. **Modern operating system**. [S.l.]: Pearson Education, Inc, 2009.

VELLAIPANDIYAN, S.; SRIKRISHNAN, V. An approach to discover the best-fit factors for the optimal performance of hadoop map reduce in virtualized environment. In: IEEE. **2014 IEEE International Conference on Computational Intelligence and Computing Research**. [S.l.], 2014. p. 1–5.

VOGEL, A.; GRIEBLER, D.; MARON, C. A.; SCHEPKE, C.; FERNANDES, L. G. Private iaas clouds: a comparative analysis of opennebula, cloudstack and openstack. In: IEEE. **2016 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP)**. [S.l.], 2016. p. 672–679.

WANG, Y.; KUNG, L.; BYRD, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. **Technological Forecasting and Social Change**, Elsevier, v. 126, p. 3–13, 2018.

WINDOWS. **Start Windows Reliability and Performance Monitor in a specific standalone mode.** [S.l.], 2018. Disponível em: <https://docs.microsoft.com/en-us/windows-server/administration/windows-commands/perfmon>. Acesso em: 19 nov. 2017.

WINSTON, W. **Microsoft Excel data analysis and business modeling**. [S.l.]: Microsoft press, 2016.

YADAV, R.; SOUSA, E.; CALLOU, G. Performance comparison between virtual machines and docker containers. **IEEE Latin America Transactions**, IEEE, v. 16, n. 8, p. 2282–2288, 2018.

YADAV, R. R.; CAMPOS, G. A.; SOUSA, E. T. G.; LINS, F. A. A strategy for performance evaluation and modeling of cloud computing services. **Revista de Informática Teórica e Aplicada**, v. 26, n. 1, p. 78–90.

YANG, C.; YU, M.; HU, F.; JIANG, Y.; LI, Y. Utilizing cloud computing to address big geospatial data challenges. **Computers, environment and urban systems**, Elsevier, v. 61, p. 120–128, 2017.

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

ZHANG, F.; SAKR, M. Performance variations in resource scaling for mapreduce applications on private and public clouds. In: IEEE. **2014 IEEE 7th International Conference on Cloud Computing**. [S.l.], 2014. p. 456–465.

APÊNDICE

APÊNDICE A - Captura de Dados do *Twitter* no R *software*

O R *software*(R, 2018) é uma aplicação que consegue capturar dados do *Twitter* com o auxílio de um pacote chamado twitteR (JEFF, 2016). Capturas de *posts* desta rede social foram realizadas de acordo com parâmetros tais como localização do comentário, idioma , polaridade do comentário, dentre outros. As linhas 1, 2, 3 e 4 instalam e requisitam o pacote necessário para acesso à dados da rede social. As linhas 5 até 9 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 10 indica quais termos devem ser parte de comentários da rede social, que neste caso são *hashtags* relacionados aos candidatos das eleições. A linha 11 armazena na variável *tweets* uma busca de 10.000 *posts* da rede social na língua portuguesa. A linha 12 tem a função de estruturar os dados de comentários da rede social para processamento no ambiente *Hadoop*. Por fim, a linha 13 salva o conjunto de dados em um arquivo no ***HDFS***.

Após a captura do *data set* relativos à comentários da rede social a partir de 28 de setembro em 2018 até 28 de outubro de 2018, e estes foram manipulados para formar cargas de trabalho definidas no planejamento de experimentos(4, 7 e 10 *GB*). O Algoritmo 1 proposto para a captura dos dados do *Twitter* é apresentado a seguir.

Algoritmo 1: Captura de dados da rede social no R *software*

```
1:  install.packages("twitteR")
2:  library(twitteR)
3:  require(twitteR)
4:  require(RCurl)
5:  api_key <- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
6:  api_secret <-"xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
7:  access_token <- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
8:  access_secret <- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
9:  setup_twitter_oauth(api_key, api_secret, access_token, access_secret)
10: terms <- c("termos a serem pesquisados")
11: tweets <- searchTwitteR(terms, n=10000,lang="pt")
12: tweets_df <- twListToDF(tweets)
13: write.csv(tweets_df, file="localização do arquivo")
```

APÊNDICE B - Algoritmo Adotado Para Calcular Termos Mais Citados no *Twitter*

Para realizar a análise *Big Data*, é preciso que instruções sejam enviadas ao *master node* do *Hadoop cluster* na nuvem privada. No estudo de caso(Capítulo 6), a ferramenta Pig Latin (GATES; DAI, 2016) foi adotada para o processamento do *dataset* gerado por comentários do Twitter. A linha 1 armazena o *dataset* na memória do *master node* do *Hadoop cluster*, utilizando como separador dos campos a vírgula. As linhas 2 e 3 guardam os campos de interesse do *dataset* e formam o *bag* de palavras, como chamado em ambientes de análise *Big Data*. A linha 4 agrupa as mesmas palavras e a linha 5 realiza a contagem de palavras. Na linha 6, tem-se a ordenação de todas as palavras citadas no conjunto de dados de forma descrecente, ou seja, do termo mais citado ao menos citado. A linha 7 armazena os resultados em um arquivo no armazenamento distribuído do Hadoop.

Algoritmo 2: Algoritmo adotado para calcular palavras mais citadas na rede social

```
1: twitter1 = LOAD '/user/twitter_dataset.csv' USING PigStorage(',') AS
   (id:chararray,text:chararray,a:chararray,b:chararray,c:chararray,d:chararray;)
2: select_col = FOREACH twitter1 GENERATE id,text;
3: W=FOREACH select_col GENERATE FLATTEN(TOKENIZE($1)) as word;
4: grouped = GROUP words BY word;
5: wordcount = FOREACH grouped GENERATE group, COUNT(W);
6: most_cited_words = ORDER wordcount BY $1 DESC;
7: STORE cited_words INTO 'arquivo no HDFS';
```

APÊNDICE C - Script de Monitoramento Para Medição de Utilização de Processador dos *Data Nodes* Configurados em Máquinas Virtuais da Nuvem Privada

```
#!/bin/bash

# Escreve o cabecalho de identificacao dos dados
echo "%usr %sys %idle date time elapsed_time" >> logcpu.txt
echo "%usr %sys %idle date time elapsed_time"
count=0;
while [ True ]
do

   # Armazena somente os campos de interesse
   cpu= mpstat 1 1| grep all | head -n1 | awk  {print $4,$13}

   # Obtem o tempo atual, no formato RFC3339: AAAA-MM-DD HH:MM:SS
   # O tempo corresponde ao retorno do mpstat,
   # portanto  o uso de cpu eh a media do ultimo minuto
   tempo= date --rfc-3339=seconds


   # Separa data e hora do tempo obtido
   data= echo $tempo | cut -d\  -f1
   hora= echo $tempo | cut -d\  -f2| awk  BEGIN{FS="-"}{print $1}

   # Mostre na tela as inforamcoes capturadas pelos script
   echo $cpu $data $hora $count

   # Escreve no arquivo as informacoes do disco
   echo $cpu $data $hora $count>> logcpu.txt

   #Executa o script a cada X unidade de tempo
   #Comentado porque o mpstat ja espera os 60 segundos
   #sleep 60
   count= expr $count + 1

done
```

APÊNDICE D - Script de Monitoramento Para Medição de Utilização de Memória dos *Data Nodes* Configurados em Máquinas Virtuais da Nuvem Privada

```
#!/bin/bash

# Escreve o cabe alho de identificacao dos dados
echo "#Mem_total Mem_used Mem_free Mem_buffers Mem_cached Swap_used
    Swap_free Dte Time Elapsed_time" >> logmemoria.txt

echo "#Mem_total Mem_used Mem_free Mem_buffers Mem_cached Swap_used
    Swap_free Dte Time Elapsed_time"

count=0;
while [ True ]
do

# Obtem o tempo atual, no formato RFC3339: AAAA-MM-DD HH:MM:SS
    tempo= date --rfc-3339=seconds
    memory= free | grep Mem:
    swap= free | grep Swap:

# Armazena somente os campos de interesse: Free, Used, Buffers,
    Cached, ...
    memtotal= echo $memory | awk  {print $2}
    memused= echo $memory | awk  {print $3}
    memfree= echo $memory | awk  {print $4}
    membuff= echo $memory | awk  {print $6}
    memcache= echo $memory | awk  {print $7}

    swapfree= echo $swap | awk  {print $4}
    swapused= echo $swap | awk  {print $3}

    # Separa data e hora do tempo obtido
    data= echo $tempo | cut -d\  -f1
    hora= echo $tempo | cut -d\  -f2| awk  BEGIN{FS="-"}{print $1}

    # Mostra na tela as informacoes capturadas pelos script
```

```
32    echo $memtotal $memused  $memfree   $membuff  $memcache
          $swapused  $swapfree $data $hora $count
33
34    # Escreve no arquivo as informacoes da memoria
35    echo $memtotal $memused  $memfree   $membuff  $memcache
          $swapused  $swapfree $data $hora $count >> logmemoria.txt
36
37    count = expr $count + 1
38    # Executa o script a cada X unidade de tempo
39    sleep 1
40    done
```

APÊNDICE E - *Script* para coleta de consumo de energia (SILVA, 2017) na plataforma Arduino Create (ARDUINO, 2018)

```
#include "EmonLib.h"
#include <LiquidCrystal.h>

EnergyMonitor emon1;
LiquidCrystal lcd(12, 11, 5, 4, 3, 2);

//Tensao da rede eletrica
int rede = 220.0;

//Pino do sensor SCT
int pino_sct = 0;

void setup()
{Serial.begin(9600);
  //Pino, calibracao - Cur Const= Ratio/BurdenR. 1800/62 = 29.
  emon1.current(pino_sct, 29);
  //Informacoes iniciais display
}
void loop()
{
  //Calcula a corrente
  double Irms = emon1.calcIrms(1480);
  //Mostra o valor da corrente
  Serial.print("Corrente : ");
  Serial.print(Irms-0.06); // Irms

  //Calcula e mostra o valor da potencia (Watts)
  Serial.print(" Potencia : ");
  Serial.println((Irms-0.06)*rede);

  delay(1000);
}
```