Predição de Falhas em Ambientes de Big Data: Revisão Sistemática de Literatura

Gustavo Hammerschmidt
PUCPR - Pontíficia Universidade Católica Paraná
Ciência da Computação
Email: g.hammerschmidt@pucpr.edu.br

O número de informações geradas por usuários na internet duplica em tamanho a cada dois anos. Esse aumento fomenta o uso de ferramentas de big data, que se popularizam mais com o avanço das tecnologias de ambiente e computação em nuvem. Soluções em big data são requisitadas para análise de dados com o ímpeto de prover inteligência comercial. Porém, para prover todo esse poder de pesquisa, centros de big data enfrentam dificuldades estruturais devido a escalabilidade de seus recursos. Esses centros necessitam acompanhamento contínuo, pois processam enormes volumes de dados que, por vezes, falham em seu carregamento, se perdem, sofrem alterações, etc. Algoritmos de previsão de falhas de ambiente foram criados com o intuito de prover maior controle do ambiente e auxiliar no desempenho deste. Esta revisão sistemática de literatura tratará de como esses algoritmos ajudam nos ambientes de big data, as falhas mais recorrentes, e o atual estado da arte.

Keywords— big, data, environment, fault, prediction, revisão, sistemática, literatura, rsl

1 Introdução

Ambientes de Big Data possuem grande poder de processamento de dados, devido ao tamanho dos dados que precisam de tratamento. Para tal cenário, o uso de uma infraestrutura de computadores interconectados é o modo pelo qual cientistas de dados operam grandes volumes de informações. Contudo, tais ambientes aumentam a propensão dos computadores a falhas, levando a perdas de partes de arquivos, inatividade e lentidão. Alguns algoritmos de previsão de falhas de ambiente foram criadas, em decorrência disso, para detectar possíveis falhas e resolvê-las antes de afetar o ambiente como um todo.

A detecção de falhas durante execução ganhou muita atenção com o aumento dos dados gerados[SS], os primeiros algoritmos utilizavam-se de muitas técnicas de classificação de suscetibilidade a falhas; muitas destas generalizam muito situações específicas de ambientes, reduzindo a eficiência do algoritmo em conduzir parâmetros de rotina adequados.

Entretanto, algumas empresas ainda não produzem um vasto conjunto de dados, e, para aplicar algoritmos de análise, utilizam-se de conjuntos públicos ou de empresas com ambientes similares, o que pode levá-las a conclusões precipitadas devido à natureza do algoritmo utilizado, que performa melhor com dados condizentes ao ambiente a que é aplicado. Neste artigo, apresentaremos os resultados de nossa revisão sistemática de literatura sobre as análises feitas para prever falhas e manter o ambiente em execução, os métodos utilizados por nós para a condução da pesquisa e os critérios de avaliação dos estudos escolhidos.

2 Trabalho relacionado

Esta revisão sistemática de literatura(RSL) foi inspirada nos trabalhos de Patrick Mikalef[P+] e Cigdem Avci[CBI]. Os artigos[EDS][SS][Dai+][Son+][Vu+] utilizados como suporte à pesquisa explicitam a aplicação de algoritmos de predição de falha a ambientes de big data. Segundo Qingquan Song, et al[Son+], há muito território a ser analisado por estes algoritmos de análise de forma a classificar os problemas de escalabilidade desses ambientes.

Tabela 1. Tabela 1: Inclusão e Exclusão dos artigos selecionados

	I1	E1	E2	E3	Final
SpringerLink	508.537	20	14	2	2
ACM	49.571	20	8	3	3

Alguns algoritmos utilzam técnicas diversas na resolução de vários problemas, generalizando as anomalias encontras com base em características obtidas de outros ambientes, podendo impactar negativamente o desempenho de alguns ambientes. Contudo, o desenvolvimento de algoritmos de predição ainda é limitado[SS] a teorias utilizadas por empresas de tecnologia em seus domínios, limitando o uso de seus algoritmos a padrões ou rotinas encontradas em seus ambientes. Para que esses algoritmos consigam derivar informações relevantes, é importante entender os meios utilizados em análises de big data para estudar as falhas encontradas, desenvolver um algoritmo de aprendizado orientado a especificações, e como utilizá-lo para agregar valor ao ambiente e torná-lo mais estável.

Nós desenvolvemos uma revisão sistemática de literatura que aumeja identificar estes processos de forma a possibilitar análises mais acuradas e voltadas ao ambiente a que são aplicados; discutindo os desafios desses cenários e apontando arquiteturas para solucioná-los.

3 Método Científico

O método utilizado foi baseado nas diretrizes propostas por Kitchenham[Kit]. A nossa revisão baseia-se em artigos postados nas bases SpringerLink e ACM. A seleção dos artigos fora feita:

1) com base em artigos relacionados as palavras-chave: big data, environment, fault e prediction;

2) foram selecionados os 20 artigos mais relevantes de acordo com as bases;

3) todos os artigos duplicados foram descartados;

4) aqueles que não escritos na língua inglesa foram removidos;

5) só foram selecionados artigos postados entre 2019 e 2020.

Como apontado acima na Tabela 1, os resultados de inclusão(I1) em ambas bases foram obtidos através da busca das palavras-chave limitadas ao período 2019-2020, onde apenas artigos de língua inglesa foram selecionados. Na primeira fase de exclusão(E1), apenas 20 artigos de cada base foram selecionados com base no número de menções feitas e relevância indicados pela a respectiva base.

Na segunda fase de exclusão(E2), os artigos foram avaliados e removidos com base na duplicação entre bases e na falta de palavras-chave condizentes com a proposta de RSL. Na última fase de exclusão(E3), os artigos foram lidos e avaliados com o grau de aferição de valor à pesquisa, restando apenas 5 artigos no total.

As strings de busca utilizadas nas bases ACM e SpringerLink foram, respectivamente, [All: big data environment fault prediction] AND [Publication Date: (01/01/2019 TO 08/31/2020)] e '(big OR data OR enviroment OR OR fault OR prediction OR analysis)'; com as devidas modificações das esclusões durante a filtragem de artigos nas próprias páginas de resultados das bases.

4 Planejamento

Nosso planejamento foi baseado em um plano de execução de uma semana para esta pesquisa, onde nós dedicamos 3 dias para pesquisa de temas e abstraimos a escolha. No quarto dia, fizemos a busca dos artigos nas bases e nos focamos em selecionar e filtrar a coleção de artigos obtida. Terminando a semana com a escrita deste artigo e a formulação das ideias.

As questões de pesquisa abordadas na revisão foram:

Q1) se o algoritmo é capaz de descrever a situação de falha a que se aplica ou como a evita;

Q2) se o grau de complexidade de complexidade do ambiente de big data impacta na complexidade do algoritmo e se influencia no seu desempenho;

Q3) se as falhas de um ambiente, cujo o algoritmo foi desenvolvido, têm impacto na avaliação deste se é aplicado em um outro ambiente.

5 Condução

A condução da pesquisa tange os seguintes tópicos:

1) definição do tema;

2) buscas nas bases de períodicos, seleção e filtragem dos resultados;

3) escolha da abordagem e escrita da RSL;

4) coleta de referências e formulação do artigo. Sendo a condução divida em um período de 7 dias, como mencionado na seção Planejamento.

6 Síntese dos dados

Os dados obtidos nesta pesquisa, cujo processo de coleta foi explicitado na seção Método Científico, seguiram um processo de abstração de tópicos nos artigos avaliados, sendo assim feito: a introdução e a con-

clusão de cada artigo foram lidas previamente ante o excerto, que foi mais profundamente avaliado durante a comparação de abordagens discutidas relacionadas à previsão de falhas em ambientes de big data.

Uma comparação entre as ideias de âmbito geral foi feita para que um resultado condizente com a pesquisa fosse obtido. Especificações de alguns artigos a respeito de peculiaridades por eles tratadas foram desconsideradas. Focamos na obtenção de técnicas utilizadas para a construção de algoritmos de predição mais acurados e focados ao ambiente a que se aplicam.

7 Discussão de Resultados

Nós observamos que algoritmos de predição de possíveis falhas estão, intimamente, integrados aos sistemas que operam sobre, logo sendo bastante diferentes entre si. Contudo, nossa pesquisa apontou que abordagem por eles utilizadas pode ser abstraída a um padrão, ou uma lista de fatores que devem ser levados em consideração por algoritmos de predição, de forma a serem capazes de prover análises acuradas. Como apontado no artigo[EDS] a respeito do monitoramento de indústrias smart, o constante monitoramento dos equipamentos de ambientes de trabalho da indústria e suas informações foram fundamentais para que um estudo de falhas fosse feito. Neste estudo, o consumo de energia de alguns aparelhos sofria oscilações em momentos de falha, levando os pesquisadores a encontrar disparidades entre o desempenho esperado e o real desempenho do ambiente de trabalho. Com a coleta destas anormalidades de consumo, um conjunto indicativo de dados foi derivado para mapeamento de falhas, notificando os profissionais responsáveis sobre irregularidades. Desta forma, respondendo à questão de pesquisa 1(Q1), indicando que há impacto, porém, ele sofre influencia também do conhecimento dos analistas.

Neste estudo[SS], contudo, Santosh indica que muitos ambientes utilizam conjuntos de dados públicos para previsão de possíveis falhas em seus ambientes e algoritmos de análise de baixa performance; segundo ele, técnicas estatísticas conseguem ter um bom desempenho em classificações binárias de classes ou tipos de falhas, de forma a maximar a qualidade de análises. Nestes estudos, a resolução das falhas é feita com a maior suscetibilidade de ambientes a falhas, sendo depois reportadas e analisadas, definidas em conjunto de informações sobre o ambiente e utilizadas por um algoritmo de aprendizado para identificação de padrões, que faz análises do status de ambientes em tempo real.

Portanto, para que um ambiente seja anti-falha, é necessário que a falha tenha sido reportada ao menos uma vez e estudada; prova-se necessário um registro de todas as ações executadas em um ambiente, para futuras análises. Para Santosh, com muitos estudos, é possível derivar um algoritmo capaz de auxiliar organizações e empresas com um histórico insuficiente de falhas, mas, para isso, seria necessário um algoritmo compatível com diferentes projetos e que houvessem similaridades entre os ambientes. Respondendo, assim, a nossa questão de pesquisa 2(Q2), onde a complexidade do ambiente em que o algoritmo é desenvolvido pode proporcionar uma inflexibilidade deste a novos contextos.

Algumas análises feitas com tensors em ambientes de big data, porém, mostram que há muitas falhas a serem classificadas por estes algoritmos. Segundo Qingquan Song, et al[Son+], há muito território a ser analisado por estes algoritmos de análise de forma a classificar os problemas de escalabilidade; pois, a complexidade de administração das infraestruturas de big data e o tempo de relato limitam aplicações de práticas de tensores usados nesses algoritmos, e estes limitam-se a problemas de baixa complexidade ou rápida realização. Alguns métodos hierárquicos são aplicados a problemas de alta complexidade, porém, eles são inflexíveis e generalizam os problemas com base nas características do ambiente a que foram treinados; e, algumas estratégias podem ser inadequadas, impactando negativamente na efetividade de predição do algoritmo.

Como fora mencionada, algumas empresas são incapazes de gerar um vasto histórico de falhas de ambiente, portanto, usam algoritmos de organizações que possuam ambientes similares; porém, as análises feitas por um algoritmo são teóricas, baseadas em pressupostos estatísticos – que não necessariamente deduzem uma situação adequadamente. Como indicado na questão de pesquisa 3(Q3), as falhas de ambientes genéricas pode impacatar negativamente no desempenho de um outro ambiente. Além disso, os estudos citados ressaltam a importância de gerar dados heterogêneos para uma maior conformidade dos algoritmos a diferentes ambientes ou situações; para Qingquan Song, et al[Son+], a mescla de variedade dos dados de aplicações voltadas a big data e conhecimento do domínio tratado proporciona maiores conhecimentos do ambiente e direciona os algoritmos a fazerem predições mais acuradas; outrossim, eles acreditam que a expertise dos profissionais é tão relevante quanto a simples automatização da rotina, instigando uma constante análise do ambiente pelos seus profissionais também.

8 Conclusão

Os resultados obtidos indicam que ambientes precisam sofrer falhas repetidas vezes para que um conjunto de dados seja derivado, e possa proporcionar acurácia ao algoritmo de predição; é necessário, também, um conhecimento das especificações dos dispositivos do ambiente para que um relatório do funcionamento ideal desses seja formulado e utilizado por algoritmos de aprendizado como referência, Todo o histórico do ambiente gerado é utilizado para a construção de algoritmos de aprendizado próprio, que auferem análises ao ambiente, reportando anormalidades conhecidas aos profissionais responsáveis. Estes, então, definem comportamentos de rotina na ocorrência das anormalidades para automatizar a execução do ambiente. É também importante para o ambiente que haja a integração dos resultados de predição e o uso da expertise de seus profissionais para proporcionar um melhor funcionamento do ambiente como um todo.

Referências

[CBI] Avci C., Tekinerdogan B. e Athanasiadis I.N. "Software architectures for big data: a systematic literature review." Em: *Big Data Anal 5, 5 (2020).* (). DOI: https://doi.org.ez433.periodicos.capes.gov.br/10.1186/s41044-020-00045-1.

[Dai+] Hong-Ning Dai et al. "Big Data Analytics for Large-scale Wireless Networks: Challenges and Opportunities." Em: *ACM Comput. Surv. 52, 5, Article 99 (September 2019), 36 pages.* (). DOI: https://doi.org/10.1145/3337065.

[EDS] Kim E., Huh D. e Kim S. "Knowledge-based power monitoring and fault prediction system for smart factories." Em: *Pers Ubiquit Comput (2019).* (). DOI: https://doi.org.ez433.periodicos.capes.gov.br/10.1007/s00779-019-01348-4.

[Kit] Barbara Kitchenham. "Procedures for Performing Systematic Reviews." Em: *NICTA Technical Report 0400011T.1* (). DOI: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf.

[P+] Mikalef P. et al. "Big data analytics capabilities: a systematic literature review and research agenda." Em: *Inf Syst E-Bus Manage 16, 547–578 (2018).* (). DOI: https://doi.org.ez433.periodicos.capes.gov.br/10.1007/s10257-017-0362-y.

[SS] Rathore S.S. e Kumar S. "A study on software fault prediction techniques." Em: *Artif Intell Rev 51, 255–327 (2019).* (). DOI: https://doi.org.ez433.periodicos.capes.gov.br/10.1007/s10462-017-9563-5.

[Son+] Qingquan Song et al. "Tensor Completion Algorithms in Big Data Analytics." Em: *ACM Trans. Knowl. Discov. Data 13, 1, Article 6 (January 2019), 48 pages.* (). DOI: https://doi.org/10.1145/3278607.

[Vu+] Tin Vu et al. "Using Deep Learning for Big Spatial Data Partitioning." Em: *ACM Trans. Spatial Algorithms Syst. 7, 1, Article 3 (August 2020), 37 pages.* (). DOI: https://doi.org/10.1145/3402126.