

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**UMA IMPLEMENTAÇÃO PARALELA DO
ALGORITMO DE DETECÇÃO DE NOVIDADE
EM STREAMS MINAS**

LUÍS HENRIQUE PUHL DE SOUZA

ORIENTADOR: PROF. DR. HERMES SENGER

São Carlos – SP

Fevereiro/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**UMA IMPLEMENTAÇÃO PARALELA DO
ALGORITMO DE DETECÇÃO DE NOVIDADE
EM STREAMS MINAS**

LUÍS HENRIQUE PUHL DE SOUZA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Sistemas de Computação

Orientador: Prof. Dr. Hermes Senger

São Carlos – SP

Fevereiro/2020

AGRADECIMENTOS

O presente trabalho de pesquisa está sendo realizado com o apoio parcial do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

LISTA DE FIGURAS

2.1	Paradigma de <i>Edge Computing</i> (SHI et al., 2016).	13
2.2	Arquitetura <i>Lambda</i> com detalhes práticos (KREPS, 2014).	16

Capítulo 6

CRONOGRAMA

Neste capítulo apresentam-se as etapas previstas e sua distribuição temporal.

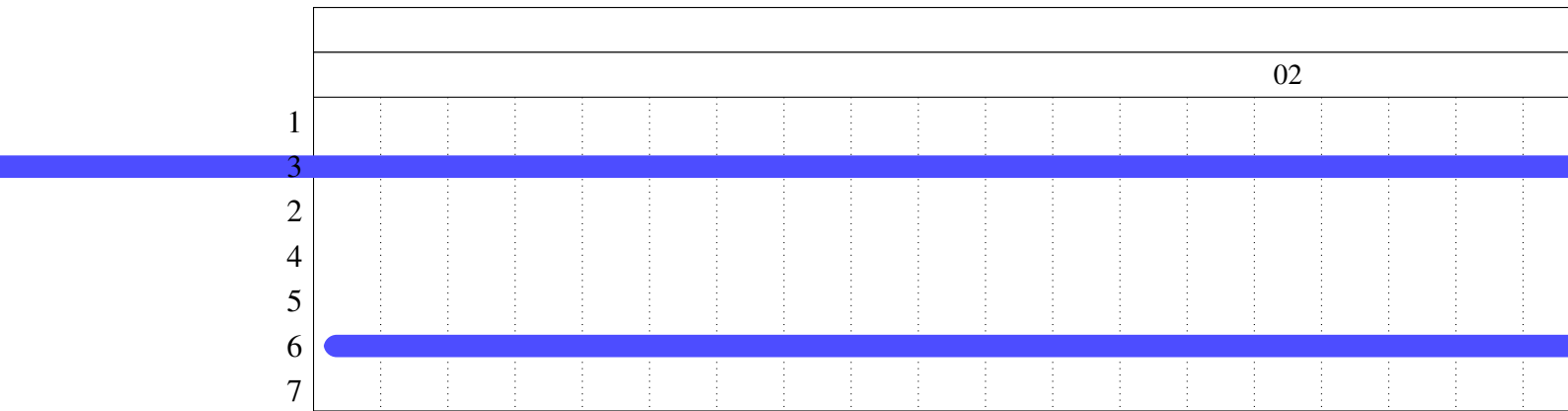
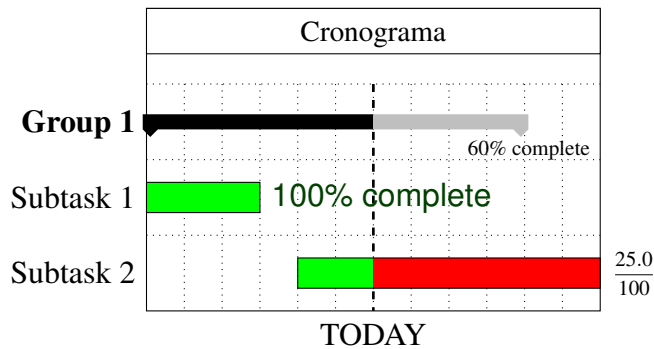
- Enumerar métricas de qualidade de classificação e métricas de escalabilidade encontradas na literatura;
- Avaliar plataformas de processamento distribuído de fluxos como:
 - *Apache Kafka* com *Python*;
 - *Apache Kafka Streams*;
 - *Apache Spark Streaming*;
 - *Apache Storm*;
 - *Apache Flink*;
- Implementar algoritmo MINAS sobre *Apache Flink*;
- Executar a implementação com *data sets* públicos relevantes;
- Validar, por meio de comparação com o algoritmo original, se a implementação gera os mesmos resultados quanto a classificação;
- Extrair métricas de escalabilidade por meio de experimentação;
- Avaliar estratégias de distribuição do algoritmo em *fog* como:
 - Detecção de novidade nas pontas;
 - Detecção de novidade na nuvem;
 - Detecção de novidade nas pontas e em nuvem;

e seu o impacto na qualidade de classificação e volume computado.

[label=)] Desenvolvimento da aplicação; Validação da aplicação em contraste com a implementação MINAS original:

2.
- preparação e, se necessário, adaptação da implementação original e *data sets*;
 - comparação e, se necessário, ajustes à implementação.
3. Experimentos com *data sets* e estratégias de distribuição em *fog*;
4. Comparação com outros trabalhos, envolvendo:
 - preparação e, se necessário, adaptação dos programas e *data sets*;
 - comparação e publicação dos resultados.
5. Preparação da base de dados para treinamento e teste: Mês 02
6. Teste ... : meses 04 a 07
7.

Quadro (linhas: etapas x colunas: meses)



REFERÊNCIAS

ABDALLAH, Z. S.; GABER, M. M.; SRINIVASAN, B.; KRISHNASWAMY, S. Anynovel: detection of novel concepts in evolving data streams. *Evolving Systems*, Springer, v. 7, n. 2, p. 73–93, 2016.

Apache Hadoop. *The ApacheTM Hadoop® project develops open-source software for reliable, scalable, distributed computing*. 2020. Disponível em: <https://hadoop.apache.org/>.

Apache Spark. *Apache SparkTM - Unified Analytics Engine for Big Data*. 2020. Disponível em: <https://spark.apache.org/>.

BIFET, A.; HOLMES, G.; KIRKBY, R.; PFAHRINGER, B. MOA: massive online analysis. *J. Mach. Learn. Res.*, v. 11, p. 1601–1604, 2010. Disponível em: <http://portal.acm.org/citation.cfm?id=1859903>.

BONOMI, F.; MILITO, R.; ZHU, J.; ADDEPALLI, S. Fog computing and its role in the internet of things. In: *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. [s.n.], 2012. p. 13–16. ISBN 9781450315197. Disponível em: <http://www.lispmob.org/>.

CASSALES, G. W.; SENGER, H.; DE FARIA, E. R.; BIFET, A. IDSA-IoT: An Intrusion Detection System Architecture for IoT Networks. In: *2019 IEEE Symposium on Computers and Communications (ISCC)*. [s.n.], 2019. p. 1–7. ISBN 978-1-7281-2999-0. ISSN 1530-1346. Disponível em: <https://ieeexplore.ieee.org/document/8969609/>.

COSTA, J. D.; FARIA, E. R.; SILVA, J. A.; GAMA, J.; CERRI, R. Novelty detection for multi-label stream classification. *Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*, n. 8, p. 144–149, 2019.

Da Silva, T. P.; SCHICK, L.; De Abreu Lopes, P.; De Arruda Camargo, H. A fuzzy multiclass novelty detector for data streams. *IEEE International Conference on Fuzzy Systems*, IEEE, v. 2018-July, p. 1–8, 2018. ISSN 10987584.

DASTJERDI, A. V.; BUYYA, R. Fog computing: Helping the internet of things realize its potential. *Computer*, IEEE, v. 49, n. 8, p. 112–116, Aug 2016. ISSN 1558-0814.

DEAN, J.; GHEMAWAT, S. MapReduce: Simplified data processing on large clusters. *OSDI 2004 - 6th Symposium on Operating Systems Design and Implementation*, p. 137–149, 2004. ISSN 23487852.

FARIA, E. R.; GONÇALVES, I. J. C. R.; CARVALHO, A. C. P. L. F. de; GAMA, J. Novelty detection in data streams. *Artificial Intelligence Review*, Springer, v. 45, n. 2, p. 235–269, Feb 2016. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-015-9444-8>.

- FARIA, E. R. d.; CARVALHO, A. C. Ponce de L. F.; GAMA, J. Minas: multiclass learning algorithm for novelty detection in data streams. *Data Mining and Knowledge Discovery*, v. 30, n. 3, p. 640–680, May 2016. ISSN 1573-756X. Disponível em: <https://doi.org/10.1007/s10618-015-0433-y>.
- FONTUGNE, R.; BORGNAT, P.; ABRY, P.; FUKUDA, K. MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking. In: *ACM CoNEXT '10*. Philadelphia, PA: [s.n.], 2010. p. 1–12.
- GABER, M. M.; ZASLAVSKY, A.; KRISHNASWAMY, S. Mining data streams: A review. *SIGMOD Rec.*, Association for Computing Machinery, New York, NY, USA, v. 34, n. 2, p. 18–26, jun. 2005. ISSN 0163-5808. Disponível em: <https://doi.org/10.1145/1083784.1083789>.
- GAMA, J.; RODRIGUES, P. P. Data stream processing. In: _____. *Learning from Data Streams: Processing Techniques in Sensor Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 25–39. ISBN 978-3-540-73679-0. Disponível em: https://doi.org/10.1007/3-540-73679-4_3.
- IEEE Communications Society. *IEEE Std 1934-2018 : IEEE Standard for Adoption of Open-Fog Reference Architecture for Fog Computing*. IEEE, 2018. 176 p. ISBN 9781504450171. Disponível em: <https://ieeexplore.ieee.org/document/8423800>.
- KAMBOURAKIS, G.; KOLIAS, C.; STAVROU, A. The Mirai botnet and the IoT Zombie Armies. In: *MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM)*. IEEE, 2017. v. 2017-Octob, p. 267–272. ISBN 978-1-5386-0595-0. Disponível em: <http://ieeexplore.ieee.org/document/8170867/>.
- KREPS, J. *Questioning the Lambda Architecture* – O'Reilly. 2014. 10 p. Disponível em: <https://www.oreilly.com/radar/questioning-the-lambda-architecture/>.
- LOPEZ, M. A. *A monitoring and threat detection system using stream processing as a virtual function for Big Data*. Tese (Theses) — Sorbonne Université ; Universidade federal do Rio de Janeiro, jun. 2018. Disponível em: <https://tel.archives-ouvertes.fr/tel-02111017>.
- MARZ, N.; WARREN, J. *Big Data: Principles and best practices of scalable real-time data systems*. [S.l.]: New York; Manning Publications Co., 2015.
- MAWI Working Group Traffic Archive. *Index of /mawi/samplepoint-F*. 2020. Disponível em: <http://mawi.wide.ad.jp/mawi/samplepoint-F/>.
- MELL, P.; GRANCE, T. The NIST definition of cloud computing: Recommendations of the National Institute of Standards and Technology. In: NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *Public Cloud Computing: Security and Privacy Guidelines*. 2012. p. 97–101. ISBN 9781620819821. Disponível em: <http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf>.
- SHI, W.; CAO, J.; ZHANG, Q.; LI, Y.; XU, L. Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal*, Institute of Electrical and Electronics Engineers Inc., v. 3, n. 5, p. 637–646, oct 2016. ISSN 23274662. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7488250>.

SILVA, T. P. da. *Abordagem Fuzzy para Detecção de Novidade em Fluxo Contínuo de Dados*. 89 p. Tese (Master) — Universidade Federal de São Carlos, 2018. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/10544>.

VIEGAS, E.; SANTIN, A.; BESSANI, A.; NEVES, N. Bigflow: Real-time and reliable anomaly-based intrusion detection for high-speed networks. *Future Generation Computer Systems*, v. 93, p. 473 – 485, 2019. ISSN 0167-739X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X18307635>.

ZAHARIA, M.; CHOWDHURY, M.; FRANKLIN, M. J.; SHENKER, S. *Spark: Cluster Computing with Working Sets*. [S.l.].