MLOps:

Potencializando a Escalabilidade de Modelos de Negócio

Heinz F. C. Rahmig

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

A figura simboliza o resultado que desejo obter com a minha pesquisa, na qual exploro a intersecção entre Machine Learning Operations (MLOps) e a escalabilidade dos modelos de negócio de forma eficiente. Cada elemento da imagem foi cuidadosamente escolhido para representar os pilares fundamentais desta interação: as redes neurais estilizadas representam a complexidade e o poder do machine learning; os engrenagens simbolizam o mecanismo de operações que torna possível a implementação eficiente desses modelos; os ícones de nuvens destacam a importância da computação em nuvem para a armazenagem e processamento de dados, um componente crucial na infraestrutura de MLOps e, por fim, as setas ascendentes refletem o crescimento e a escalabilidade que busco alcançar nos modelos de negócio por meio da integração de MLOps.

Portanto, esta escolha de representação visual é um reflexo das complexidades do meu estudo. Através dela, busco ilustrar não apenas a importância de adotar MLOps para a escalabilidade dos negócios, mas também como essa abordagem pode ser o caminho para a inovação contínua e a vantagem competitiva no mercado.

A imagem foi criada com a ajuda do DALL-E, exemplificando o uso inovador de tecnologias de inteligência artificial que minha pesquisa enfatiza.

UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA (INF)

HEINZ FELIPE CAVALCANTE RAHMIG

MLOPS: POTENCIALIZANDO A ESCALABILIDADE DE MODELOS DE NEGÓCIO

Goiânia
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): **HEINZ FELIPE CAVALCANTE RAHMIG**

Título do trabalho: **MLOPS: POTENCIALIZANDO A ESCALABILIDADE DE MODELOS DE NEGÓCIO**

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [**X**] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Heinz Felipe Cavalcante Rahmig**, **Discente**, em 17/02/2024, às 00:37, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fernando Marques Federson**, **Professor do Magistério Superior**, em 12/09/2024, às 10:59, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

HEINZ FELIPE CAVALCANTE RAHMIG

MLOPS: POTENCIALIZANDO A ESCALABILIDADE DE MODELOS DE NEGÓCIO

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.
Orientador: Prof. Dr. Fernando Marques Federson

Goiânia
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

RAHMIG, HEINZ FELIPE CAVALCANTE
MLOPS: POTENCIALIZANDO A ESCALABILIDADE DE MODELOS DE NEGÓCIO [manuscrito] / HEINZ FELIPE CAVALCANTE RAHMIG. - 2024.
75 f.

Orientador: Prof. Dr. FERNANDO MARQUES FEDERSON.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Informática (INF), Inteligência Artificial, Goiânia, 2024.

1. inteligência artificial. 2. machine learning operations. 3. escalabilidade. 4. modelos de negócio. I. FEDERSON, FERNANDO MARQUES, orient. II. Título.

CDU 004

HEINZ FELIPE CAVALCANTE RAHMIG

MLOPS: POTENCIALIZANDO A ESCALABILIDADE DE MODELOS DE NEGÓCIO

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Data da Aprovação: 08 de fevereiro de 2024.

Prof. Dr. Fernando Marques Federson
Orientador (INF-UFG)

Prof. Dr. Aldo André Díaz Salazar
Coordenador de TCC do BIA (INF-UFG)

Prof. Dr. Vinícius Sebba Patto
Coordenador do BIA (INF-UFG)

gov.br
Documento assinado digitalmente
LEONARDO AFONSO AMORIM
Data: 09/02/2024 09:41:30-0300
Verifique em https://validar.iti.gov.br

Dr. Leonardo Afonso Amorim
(CEIA-UFG)

HEINZ F. C. RAHMING

MLOPS: POTENCIALIZANDO A ESCALABILIDADE DE MODELOS DE NEGÓCIO

RESUMO

Este Relatório de Conclusão de Curso tem como objetivo reunir os resultados da minha jornada para me tornar um especialista em **MLOps (Modelos de Negócio)**. Uma ilustração e sua narrativa descrevem os períodos de trabalho. Os Apêndices contêm os Termos de Aceite de Entrega e os resultados obtidos durante cada período de trabalho.

Palavras-chave: inteligência artificial, machine learning operations, escalabilidade, modelos de negócio.

ABSTRACT

This Course Completion Report aims to bring together the results of my journey to become an expert in **MLOps (Business Models)**. An illustration and its narrative describe the work periods. The Appendices contain the Delivery Acceptance Terms and the results obtained during each work period.

Keywords: artificial intelligence, machine learning operations, scalability, business models.

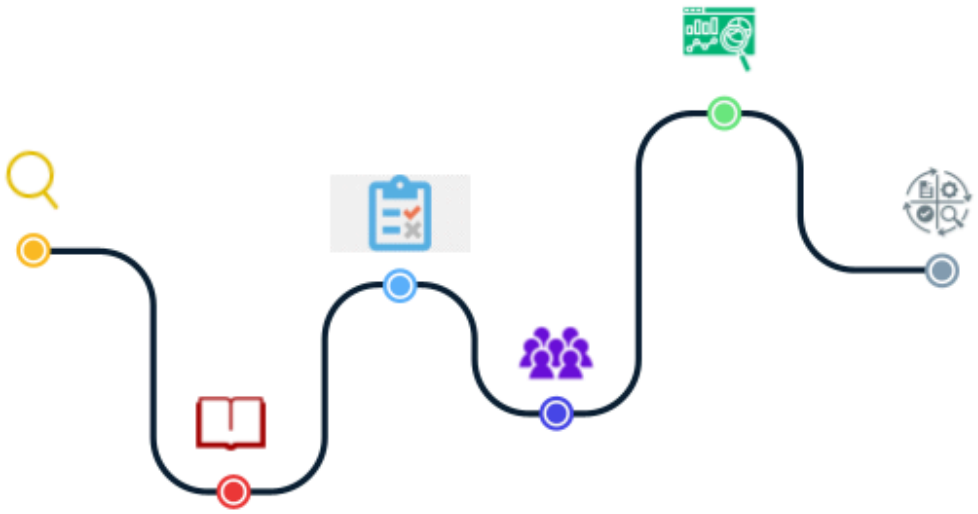
Goiânia

2024

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Minha Jornada

Heinz Felipe C. Rahmig



Especialista em: MLOps (Modelos de Negócio)

Semana 1

Definição da Área de Conhecimento para a especialização

Semanas 2 e 3

Revisão bibliográfica e leitura de artigos sobre MLOps

Semana 4

Criação de vocabulário sobre MLOps e criação de repositório de código

Semanas 5 e 6

Definição do objeto de estudo "MLOps aplicado a DSAaS" e levantamento bibliográfico sobre o modelo de negócio

Semana 7-9

Testes com frameworks de MLOps e análise de resultados obtidos.

Semana 10

Análise de Framework indicado para maturidade do negócio e considerações finais.

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

MINHA JORNADA

Nome: Heinz Felipe Cavalcante Rahmig

Especialidade: MLOps (Modelos de Negócio)

Objetivo deste documento

Durante o processo da disciplina Residência em IA[1], foram gerados diversos resultados na construção da minha especialização. A cada semana, um conjunto de resultados foi formalizado por um Termo de Aceite de Entrega e avaliado por uma banca, considerando o planejado e o realizado para o período. Este documento tem como objetivo descrever esses resultados obtidos, fazendo referência aos Termos de Aceite de Entrega e seus documentos associados.

Minha Jornada

Iniciei minha trajetória na primeira **Semana** com a tarefa de identificar o campo de estudo para minha especialização. Participei de um grupo com mais dois colegas pesquisadores e, através de reuniões iniciais, concordamos em focar em *Machine Learning Operations* (MLOps) como nossa área de interesse comum. Em seguida, iniciamos uma pesquisa por literatura científica relevante para fundamentar os trabalhos individuais de cada membro do grupo.

O primeiro passo foi explorar o acervo da conferência *Computational Science and Computational Intelligence 2023* (CSCI), onde identificamos duas áreas principais de pesquisa: Inteligência Artificial e Big Data com Ciência de Dados. No **Apêndice 1**, apresento um documento que detalha a seleção de temas nessas áreas, incluindo os termos-chave para os estudos em MLOps. Adicionalmente, há uma tabela listando publicações da CSCI de 2018 a 2023 que são pertinentes aos termos selecionados, oferecendo uma base para futuras revisões literárias.

[1] Dez semanas, entre setembro de 2023 e janeiro de 2024.

Nas **Semanas 2 e 3**, dediquei-me à revisão da literatura, examinando artigos tanto da CSCI quanto de outras fontes consideradas importantes na área de Computação. No **Apêndice 2**, há ilustrações do processo de revisão bibliográfica utilizando a ferramenta Parsifal. Este **Apêndice** detalha os objetivos da revisão, as perguntas de pesquisa, critérios de inclusão e exclusão, palavras-chave, string de busca e as bases consultadas (ACM, IEEE e Scopus). Além disso, há uma tabela com os artigos revisados pelo nosso grupo de MLOps, incluindo observações relevantes. O documento referente às atividades desenvolvidas na **Semana 3** não inclui apenas esta tabela, mas também apresenta dois elementos adicionais: um cronograma inicial para orientar as atividades dos pesquisadores e um esboço do vocabulário com os conceitos-chave identificados durante a análise dos artigos.

Na **Semana 4**, concluímos nossas atividades em grupo, focando especialmente em conceitos avançados de DevOps. Paralelamente, defini o tema da minha especialização, escolhendo investigar o papel do MLOps na promoção da escalabilidade em modelos de negócio. Com isso em mente, elaborei um documento detalhado, apresentado no **Apêndice 3**, que examina diversos modelos de negócio categorizados como *"As a Service"*. Neste documento, explorei como o MLOps pode ser integrado a esses modelos para otimizar processos, aumentar a eficiência e promover a escalabilidade sustentável.

Em face da duração da Disciplina de Residência em IA, não considerei possível realizar um estudo abrangente sobre a influência do MLOps na escalabilidade de diversos modelos de negócio. Portanto, na **Semana 5**, decidi focar em um exemplo específico: o modelo *Data Science as a Service* (DSaaS). Esta escolha se deu pela natureza inovadora e ainda não amplamente examinada do DSaaS, e pelo seu expressivo potencial de mercado, que ostenta uma taxa de crescimento anual composta de 27,7%, conforme detalhado no **Apêndice 4**. Neste **Apêndice**, descrevo minuciosamente o modelo DSaaS, investigando as características distintas do DSaaS, incluindo a procura por soluções customizadas em Ciência de Dados e a necessidade de processamento e análise de grandes volumes de dados.

Nas **Semanas 6 e 7**, dediquei-me ao desenvolvimento de um estudo sobre a integração do MLOps no modelo de negócio DSaaS, com o objetivo de identificar e abordar pontos críticos inerentes a esse modelo. A minha análise, detalhada nos **Apêndices 5 e 6**, concentrou-se em como o MLOps pode ser empregado para superar esses desafios, otimizando o modelo DSaaS para uma operação mais eficaz e escalável.

No **Apêndice 5**, apresento um diagnóstico dos principais desafios enfrentados pelo modelo DSaaS, incluindo a complexidade na gestão de projetos de Ciência de Dados, a necessidade de processamento de dados em tempo real e a importância de manter a qualidade e a segurança dos dados. Este **Apêndice** também discute a dificuldade em escalar operações de Ciência de Dados de forma eficiente, considerando o crescente volume e a complexidade dos dados.

Em seguida, no **Apêndice 6**, explorei como a implementação de práticas de MLOps pode oferecer soluções para esses pontos críticos. A análise incluiu a automação de *workflows* de Ciência de Dados, a implementação de *pipelines* de dados mais robustos e seguros, e a utilização de ferramentas de monitoramento e otimização de modelos em produção. Discuto também a importância de uma cultura de colaboração entre equipes de dados e operações, facilitando uma integração contínua e entrega contínua (CI/CD) para projetos de Ciência de Dados.

Nas **Semanas 8 e 9**, concentrei meus esforços na análise, teste e coleta de resultados utilizando frameworks específicos da AWS e do Google Cloud, com o objetivo de avaliar suas capacidades e adequação ao modelo de negócio DSaaS. Esta fase do estudo foi meticulosamente documentada nos **Apêndices 7 e 8**, onde explorei a funcionalidade, a eficiência e a escalabilidade oferecidas por cada um dos frameworks em diferentes cenários de aplicação.

No **Apêndice 7**, expus o processo de teste dos frameworks, incluindo os critérios de seleção, as configurações de teste e os cenários de uso avaliados, com o intuito de verificar como cada framework suporta operações de Ciência de Dados em larga escala, abrangendo desde a automação de processos até o gerenciamento de vastos volumes de dados, a implementação de modelos de aprendizado de máquina, e a manutenção da segurança e da

conformidade dos dados. Complementarmente, no **Apêndice 8**, concentrei-me na coleta e análise dos resultados desses testes, realizando uma comparação entre o desempenho, a usabilidade e a capacidade de integração dos frameworks da AWS e do Google Cloud conforme as demandas do DSaaS. Este estudo oferece uma análise tanto quantitativa quanto qualitativa, salientando os pontos fortes e fracos de cada framework em aspectos como facilidade de uso, flexibilidade, escalabilidade, custo e suporte disponível, fornecendo uma visão abrangente sobre suas adequações às necessidades do DSaaS.

Na **Semana 10**, procedi com a análise final para determinar qual framework seria o mais apropriado para uma empresa com o modelo de negócio DSaaS, considerando seu grau de maturidade. Essa análise final e as considerações foram compiladas no **Apêndice 9**. Neste documento, sintetizei os *insights* obtidos durante as fases de teste e avaliação, ponderando sobre como a escolha do framework impacta a eficiência operacional, a inovação e a capacidade de atender às demandas do mercado de forma competitiva.

O **Apêndice 9** também incluiu recomendações estratégicas, levando em conta tanto as capacidades técnicas quanto os aspectos econômicos associados à implementação dos frameworks. A conclusão do estudo enfatiza a importância de selecionar um framework que não apenas atenda às necessidades atuais da empresa, mas que também ofereça a flexibilidade e escalabilidade necessárias para suportar o crescimento futuro e a evolução do modelo de negócio DSaaS.

Em resumo, acredito que o estudo desenvolvido e documentado fornece uma base sólida para decisões informadas sobre tecnologias em nuvem e MLOps, direcionando empresas de DSaaS na escolha de ferramentas que possam potencializar suas operações e estratégias de crescimento no mercado de Ciência de Dados.

APÊNDICE 1

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 19 de out. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Állan Christoffer Pereira Silva
Gabriel da Mata Marques
Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Esta entrega consistiu na classificação do grupo de trabalho responsável pela temática de MLOps.

Os requisitos básicos para a entrega eram:

- Classificar os estudos segundo a metodologia da Conference on Computational Science and Computational Intelligence (CSCI).
- Pesquisar nos anais do congresso em busca de trabalhos correlatos com o tema de MLOps.

Os produtos gerados para esta entrega encontram-se nos links abaixo:

- Classificação com os termos da CSCI: https://docs.google.com/document/d/1VPub74MceGjgfJuvjHynzP3-VOxUyI8jz2pSMHIeBnY/edit?usp=sharing
- Pesquisa pelos trabalhos correlatos publicados na CSCI entre 2018 e 2022: https://docs.google.com/spreadsheets/d/1MqcCFht8JVCCPX50I3XKQAxUTjjUlg7v4px4G1jbU3M/edit?usp=sharing

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega do dia 26/10/2023, estão planejadas as seguintes atividades:

- Busca por artigos e publicações em outras bases científicas.
- Construção de um repositório com os trabalhos encontrados.
- Revisão e resumos dos trabalhos considerados mais relevantes.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Repositório do grupo de trabalho: https://github.com/AllanSilva156/mlops-residencia-ia

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Entrega Gate 19/10/2023

Introdução

Este documento faz parte da entrega referente ao gate do dia 19 de outubro de 2023. Nele está detalhada a classificação do grupo de pesquisa responsável pela temática de MLOps.

Responsáveis pela Entrega:

Állan Christoffer Pereira Silva
Gabriel da Mata Marques
Heinz Felipe Cavalcante Rahmig

Classificação segundo Conference on Computational Science and Computational Intelligence (CSCI):

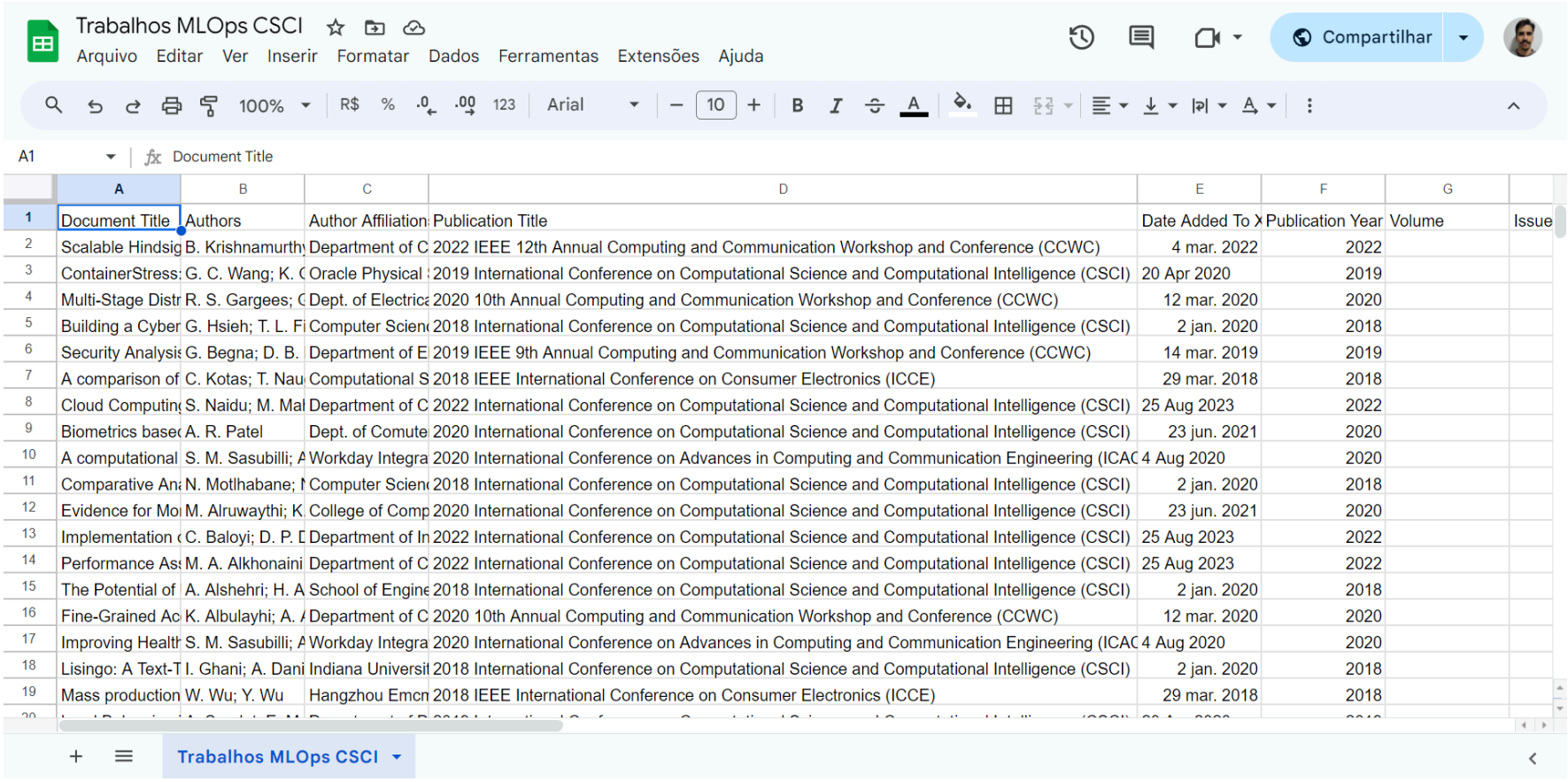
Research Track on Big Data and Data Science (CSCI-RTBD)

SECURITY & PRIVACY IN THE ERA OF DATA SCIENCE & BIG DATA:
- Privacy Preserving Big Data Collection

INFRASTRUCTURES FOR BIG DATA & DATA SCIENCE:
- Cloud Based Infrastructures (applications, storage & computing resources)
- HPC, including Parallel & Distributed Processing
- Programming Models and Environments to Support Big Data
- Software and Tools for Big Data
- Big Data Open Platforms
- Emerging Architectural Frameworks for Big Data
- Paradigms and Models for Big Data beyond Hadoop/MapReduce

BIG DATA & DATA SCIENCE MANAGEMENT AND FRAMEWORKS:
- Database and Web Applications
- Massively Parallel Processing (MPP) Databases
- Distributed Database Systems
- Distributed File Systems
- Distributed Storage Systems
- Data Preservation and Provenance
- Data Protection Methods
- Data Integrity and Privacy Standards and Policies

Revisão na *CSCI*



Trabalhos MLOps CSCI

Arquivo Editar Ver Inserir Formatar Dados Ferramentas Extensões Ajuda

Compartilhar

A1 | Document Title

	A	B	C	D	E	F	G	
1	Document Title	Authors	Author Affiliation	Publication Title	Date Added To X	Publication Year	Volume	Issue
2	Scalable Hindsig	B. Krishnamurthy	Department of C	2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)	4 mar. 2022	2022		
3	ContainerStress:	G. C. Wang; K. C	Oracle Physical	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019		
4	Multi-Stage Distr	R. S. Gargees; C	Dept. of Electrica	2020 10th Annual Computing and Communication Workshop and Conference (CCWC)	12 mar. 2020	2020		
5	Building a Cyber	G. Hsieh; T. L. Fi	Computer Scien	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018		
6	Security Analysis	G. Begna; D. B.	Department of E	2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)	14 mar. 2019	2019		
7	A comparison of	C. Kotas; T. Nau	Computational S	2018 IEEE International Conference on Consumer Electronics (ICCE)	29 mar. 2018	2018		
8	Cloud Computin	S. Naidu; M. Mal	Department of C	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022		
9	Biometrics base	A. R. Patel	Dept. of Comute	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020		
10	A computational	S. M. Sasubilli; A	Workday Integra	2020 International Conference on Advances in Computing and Communication Engineering (ICAC	4 Aug 2020	2020		
11	Comparative Ana	N. Motlhabane; M	Computer Scien	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018		
12	Evidence for Mo	M. Alruwaythi; K.	College of Comp	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020		
13	Implementation	C. Baloyi; D. P. D	Department of In	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022		
14	Performance Ass	M. A. Alkhonaini	Department of C	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022		
15	The Potential of	A. Alshehri; H. A	School of Engine	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018		
16	Fine-Grained Ac	K. Albulayhi; A. A	Department of C	2020 10th Annual Computing and Communication Workshop and Conference (CCWC)	12 mar. 2020	2020		
17	Improving Health	S. M. Sasubilli; A	Workday Integra	2020 International Conference on Advances in Computing and Communication Engineering (ICAC	4 Aug 2020	2020		
18	Lisingo: A Text-T	I. Ghani; A. Dani	Indiana Universit	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018		
19	Mass production	W. Wu; Y. Wu	Hangzhou Emcn	2018 IEEE International Conference on Consumer Electronics (ICCE)	29 mar. 2018	2018		

Trabalhos MLOps CSCI

APÊNDICE 2

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: **26 de out. de 2023**

Participantes da Entrega **[matriculados em Residência em IA]**:

Állan Christoffer Pereira Silva
Gabriel da Mata Marques
Heinz Felipe Cavalcante Rahmig

Entrega: **[descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]**

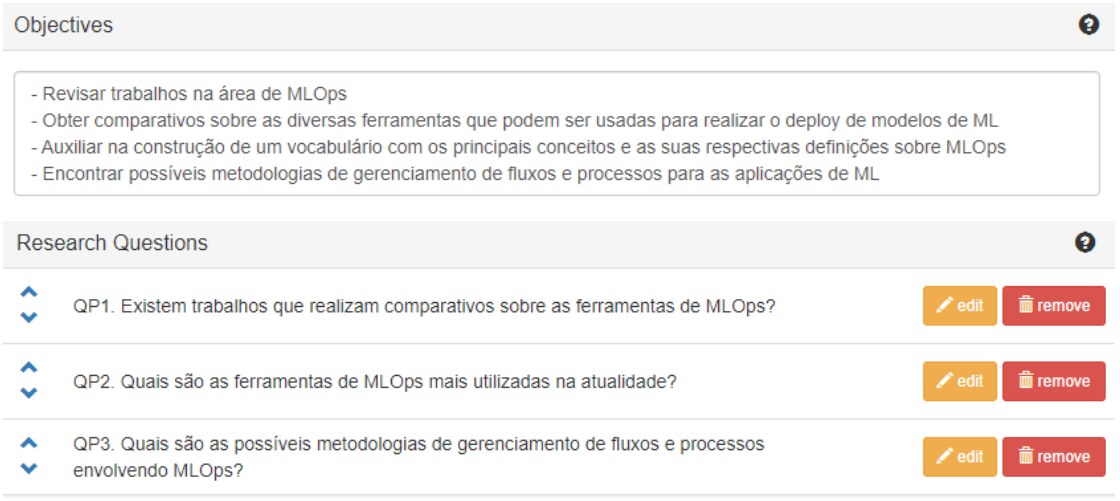
Esta entrega consistiu no início do processo de revisão bibliográfica das referências que irão fundamentar os trabalhos do grupo responsável pela temática de MLOps.

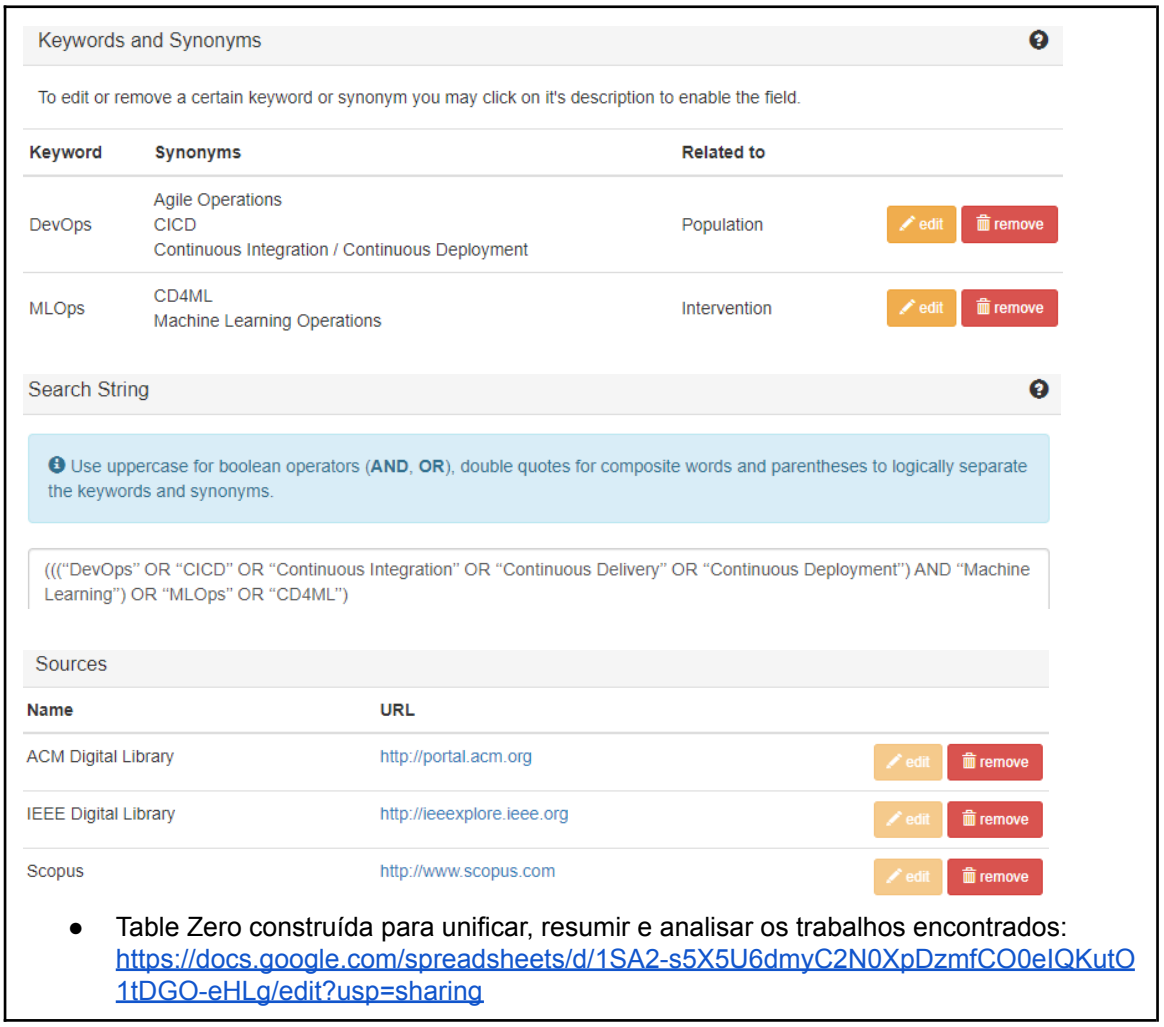
Os requisitos básicos para a entrega eram:

- Buscar artigos e publicações em outras bases científicas além da CSCI.
- Construir um repositório com os trabalhos encontrados.
- Realizar uma revisão preliminar dos trabalhos e um breve resumo sobre cada um.

Os produtos gerados para esta entrega estão descritos a seguir:

- Estruturação do processo de revisão bibliográfica utilizando a ferramenta Parsifal.





Keywords and Synonyms

To edit or remove a certain keyword or synonym you may click on it's description to enable the field.

Keyword	Synonyms	Related to	
DevOps	Agile Operations CICD Continuous Integration / Continuous Deployment	Population	edit remove
MLOps	CD4ML Machine Learning Operations	Intervention	edit remove

Search String

Use uppercase for boolean operators (**AND**, **OR**), double quotes for composite words and parentheses to logically separate the keywords and synonyms.

(((“DevOps” OR “CICD” OR “Continuous Integration” OR “Continuous Delivery” OR “Continuous Deployment”) AND “Machine Learning”) OR “MLOps” OR “CD4ML”)

Sources

Name	URL	
ACM Digital Library	http://portal.acm.org	edit remove
IEEE Digital Library	http://ieeexplore.ieee.org	edit remove
Scopus	http://www.scopus.com	edit remove

- Table Zero construída para unificar, resumir e analisar os trabalhos encontrados: https://docs.google.com/spreadsheets/d/1SA2-s5X5U6dmyC2N0XpDzmfCO0eIQKutO1tDGO-eHLg/edit?usp=sharing

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

Para a próxima entrega do dia 09/11/2023, estão planejadas as seguintes atividades:

- Finalização do processo de revisão bibliográfica.
- Construção do cronograma de atividades a serem desenvolvidas até o final da Residência.
- Início da montagem de um vocabulário com termos e definições relacionadas à temática de MLOps.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

Table Zero

Trabalhos MLOps CSCI

Arquivo Editar Ver Inserir Formatar Dados Ferramentas Extensões Ajuda

Compartilhar

A1 Document Title

	A	B	C	D	E	F	G	H	I	J	K
1	Document Title	Authors	Author Affiliation	Publication Title	Date Added To X	Publication Year	Volume	Issue	Start Page	End Page	Abstract
2	Scalable Hindsight Experience Replay based Q-l	B. Krishnamurth	Department of C	2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)	4 mar. 2022	2022			138	144	Nowadays Inter
3	ContainerStress: Autonomous Cloud-Node Scopi	G. C. Wang; K. (	Oracle Physical	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1257	1262	Deploying big-d
4	Multi-Stage Distributed Computing for Big Data: E	R. S. Gargees; (	Dept. of Electric	2020 10th Annual Computing and Communication Workshop and Conference (CCWC)	12 mar. 2020	2020			626	633	With the increas
5	Building a Cybersecurity Research and Experime	G. Hsieh; T. L. F	Computer Scien	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			30	35	Cybersecurity is
6	Security Analysis in Context-Aware Distributed SI	G. Begna; D. B.	Department of E	2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)	14 mar. 2019	2019			177	183	Recent studies
7	A comparison of Amazon Web Services and Micr	C. Kotas; T. Nau	Computational S	2018 IEEE International Conference on Consumer Electronics (ICCE)	29 mar. 2018	2018			1	4	Advances in co
8	Cloud Computing From an Architectural View	S. Naidu; M. Ma	Department of C	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			1922	1925	software design
9	Biometrics based access framework for secure cl	A. R. Patel	Dept. of Comute	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			1318	1321	This paper is fo
10	A computational and analytical approach for clou	S. M. Sasubilli; A	Workday Integra	2020 International Conference on Advances in Computing and Communication Engineering (ICA	4 Aug 2020	2020			1	5	Cloud computin
11	Comparative Analysis of Cloud Computing Simul	N. Motlhabane; I	Computer Scien	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1309	1316	Cloud computin
12	Evidence for Monitoring the User and Computing	M. Alruwaythi; K	College of Comp	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			1309	1313	Cloud computin
13	Implementation of an Enhanced Security Algorith	C. Baloyi; D. P. I	Department of Ir	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			953	957	Cloud Computir
14	Performance Assessment for Scheduling Algorith	M. A. Alkhonaini	Department of C	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			1336	1340	The concept of
15	The Potential of Utilizing Mobile Cloud Computing	A. Alshehri; H. A	School of Engine	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1328	1331	Nowadays, the
16	Fine-Grained Access Control in the Era of Cloud	K. Albulayhi; A. /	Department of C	2020 10th Annual Computing and Communication Workshop and Conference (CCWC)	12 mar. 2020	2020			748	755	Access control i
17	Improving Health Care by Help of Internet of Thin	S. M. Sasubilli; A	Workday Integra	2020 International Conference on Advances in Computing and Communication Engineering (ICA	4 Aug 2020	2020			1	4	Present days th
18	Lisingo: A Text-To-Speech Web Service-Based A	I. Ghani; A. Dani	Indiana Universi	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1323	1327	As cloud compu
19	Mass production EMC status quick evaluation thr	W. Wu; Y. Wu	Hangzhou Emcr	2018 IEEE International Conference on Consumer Electronics (ICCE)	29 mar. 2018	2018			1	4	This paper prop
20	Load Balancing in Cloud Computing Using Genet	A. Saadat; E. Ma	Department of D	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1435	1440	Cloud computin
21	Anomaly Detection in Smart Environments using	D. A. B. Moreira	State University	2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)	11 mar. 2021	2021			1	2	Modern society
22	Computing Node Selection Method Based on Pro	T. Ueda; D. Kon	Department of C	2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)	17 mar. 2023	2023			672	673	Technologies su
23	Study of Distributed Framework Hadoop and Ove	R. Solanki; S. H.	Department of C	2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)	14 mar. 2019	2019			252	257	The amount of
24	MR-Edge: a MapReduce-based Protocol for IoT E	Q. Wang; B. Lee	Software Resea	2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)	28 Feb 2019	2019			1	6	Edge computing
25	Framework modeling for User privacy in cloud co	A. Almtrf; Y. Alaç	School of Engine	2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)	14 mar. 2019	2019			819	826	Many organizati
26	Scale-out Acceleration for 3D CNN-based Lung N	J. Shen; D. Wan	College of Comp	2019 56th ACM/IEEE Design Automation Conference (DAC)	22 Aug 2019	2019			1	6	Three-dimensio
27	Real-Time Energy-Conserving VM-Provisioning F	S. Ismaeel; A. M	Department Con	2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)	14 mar. 2019	2019			765	771	This paper prop
28	A Review of Vehicular Micro-Clouds	A. Phadke; F. A.	Department of C	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			411	417	Data-intensive a
29	User Behavior Trust Modeling in Cloud Security	M. Alruwaythi; K	Department of C	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1336	1339	Evaluating user
30	Performance evaluations of multimedia service fu	K. Imagane; K. I	Graduate Schoo	2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)	19 mar. 2018	2018			1	4	As mobile multi
31	An OpenVPN-Based Interconnection in Multi-Clo	S. Jarrous-Holtr	University of Mü	2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)	17 mar. 2023	2023			867	870	Real-Time Onlir
32	Vulnerability Scanning with Google Cloud Platfor	N. J. Mitchell; K.	School of Comp	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1441	1447	Cloud is comple
33	Cloud-Based Sepsis Prediction System with Neu	Y. -H. Chiang; H	High Performan	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			19	25	Sepsis is a com
34	Efficient parallelization for big data collaborative r	E. O. Aboagye; (	UESTC, Chengc	2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)	26 Feb 2018	2018			268	274	This paper prop

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 9 de nov. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Állan Christoffer Pereira Silva
Gabriel da Mata Marques
Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Esta entrega consistiu na finalização do processo de revisão bibliográfica, o qual tinha por objetivo encontrar trabalhos de referência na área de MLOps.

Os requisitos básicos para a entrega eram:

- Finalizar o processo de revisão bibliográfica.
- Construir o cronograma de atividades a serem desenvolvidas até o final da Residência.
- Iniciar a montagem de um vocabulário com termos e definições relacionadas à temática de MLOps.

Os produtos gerados para esta entrega estão descritos a seguir:

- Repositório com os artigos encontrados e suas respectivas análises (Table Zero): https://docs.google.com/spreadsheets/d/1SA2-s5X5U6dmyC2N0XpDzmfCO0eIQKutO1tDGO-eHLg/edit?usp=sharing
- Cronograma de atividades da Residência: https://docs.google.com/spreadsheets/d/16sy4Z3gcDNV2U5fZOC_mPgE7eyiGfgc5zHQFBCrxBSc/edit?usp=sharing
- Vocabulário sobre MLOps: https://docs.google.com/document/d/1AtmA9GgHnD4mIF-bbjjbt53QLZaQAjowFyjtjQHzzLA/edit?usp=sharing

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega do dia 16/11/2023, estão planejadas as seguintes atividades:

- Finalização do vocabulário incluindo conceitos de DevOps.
- Busca por repositórios de códigos úteis.
- Decisão sobre a aplicação e levantamento de requisitos.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Cronograma de Atividades

Cronograma de Atividades

Arquivo Editar Ver Inserir Formatar Dados Ferramentas Extensões Ajuda

Compartilhar

B22

	A	B	Semanas															
	Etapa	**Descrição**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	**Objetivo principal**	Realizar estudo comparativo de implementação (dificuldades e custos) e resultados (métricas) entre modelos de ML e LLMs em tarefas de classificação de dados tabulares utilizando conceitos de DevOps e MLOps																
5	26/09/2023	Recepção e Instruções iniciais.	■															
6	05/10/2023	Planejamento.		■														
7	Gate 19/10/2023	Classificação com os termos da CSCI. Pesquisa pelos trabalhos correlatos publicados na CSCI entre 2018 e 2022.			■													
8	Gate 26/10/2023	Busca por artigos e publicações em outras bases científicas. Construção de um repositório com os trabalhos encontrados. Revisão e resumos dos trabalhos considerados mais relevantes.				■	■											
9	Gate 09/11/2023	Finalização do processo de revisão bibliográfica. Construção do cronograma de atividades a serem desenvolvidas até o final da Residência. Início da montagem de um vocabulário com termos e definições relacionadas à temática de MLOps.						■	■									
10	Gate 16/11/2023	Finalização do vocabulário incluindo conceitos DevOps. Busca por repositórios de códigos úteis. Decisão sobre a aplicação e levantamento de requisitos.								■								
11	Gate 23/11/2023	Coleta de dados. Análise exploratória.									■							
12	Gate 30/11/2023	Pré-processamento dos dados.										■						
13	Gate 07/12/2023	Construção dos modelos preditivos. LLM de baixa escala (13B) e LLM de alta escala (API OpenAI). Comparativo entre resultados do modelo de ML e LLMs.											■					
14	Gate 14/12/2023	Preparo para o deploy (requirements, prompts, Docker, etc). Elaboração da arquitetura de modelos.												■				
15	Gate 21/12/2023	Deploy de modelo de ML. Deploy de LLMs.													■			
16	Gate 11/01/2024	Ajustes e validação dos resultados. Consolidação dos resultados dos trabalhos individuais.														■		
17	15/01/2024	Elaboração do TCC.															■	
18	22/01/2024	Elaboração do TCC e apresentação.																■

Vocabulário sobre MLOps

Introdução

Este documento tem como objetivo principal a documentação dos principais termos e definições relacionados à área de Machine Learning Operations (MLOps). O conteúdo contido neste trabalho serve como base teórica para estudantes e profissionais que desejam se aprofundar sobre como operacionalizar modelos de Machine Learning, isto é, realizar a implantação de modelos preditivos.

Development Operations (DevOps)

DevOps é uma combinação de filosofias, práticas e ferramentas que aumenta a capacidade de uma organização de entregar aplicações e serviços em alta velocidade. Melhorando e evoluindo produtos mais rapidamente do que organizações que utilizam processos tradicionais de desenvolvimento e gerenciamento de infraestrutura. DevOps é caracterizado pela automação e monitoramento em todas as fases do desenvolvimento de software, desde a integração, testes, liberação até a implantação e gestão de infraestrutura.

Práticas Comuns de DevOps:

1. Integração Contínua (CI): Prática que incentiva desenvolvedores a integrar código em um repositório compartilhado. Cada check-in é então verificado por uma build automatizada, permitindo que equipes detectem problemas cedo.
2. Entrega Contínua (CD): Extensão da integração contínua para garantir que o código seja seguro e que possa ser liberado a qualquer momento.
3. Monitoramento e Logging: Processos que envolvem a coleta de métricas e logs para acompanhar o desempenho das aplicações e da infraestrutura.
4. Comunicação e Colaboração: Ferramentas e práticas culturais que promovem a colaboração dentro e entre as equipes.
5. Automação de Infraestrutura: Gerenciamento e provisionamento de infraestrutura através de código e ferramentas, minimizando a intervenção manual.

Ferramentas de DevOps:

1. Docker: Ferramenta de contêinerização que permite empacotar uma aplicação com todas as suas dependências em um contêiner padronizado.
2. Jenkins: Servidor de automação open source usado para CI/CD.
3. Kubernetes: Sistema de orquestração de contêineres que gerencia aplicações construídas em contêineres.
4. Ansible/Terraform: Ferramentas que permitem aos desenvolvedores provisionar e gerenciar infraestrutura através de código.
5. Git: Sistema de controle de versão distribuído para rastrear mudanças no código fonte durante o desenvolvimento de software.
6. Nagios/Grafana: Ferramentas de monitoramento que oferecem visibilidade em tempo real sobre a saúde da infraestrutura e aplicações.

Termos Chave de DevOps:

Termo	Definição
1- Pipeline de Deploy	Sequência de passos para entregar uma nova versão de software.
2- Infrastructure as Code (IaC)	Prática de gerenciar e provisionar infraestrutura de TI através de scripts de código.
3- Micro Serviços	Arquitetura que estrutura uma aplicação como uma coleção de serviços que são executados de forma independente.
4- Orquestração de Containers	Processo de gerenciar a vida útil de contêineres, especialmente em ambientes com muitos contêineres.
5- Gerenciamento de Configuração	Processo de manter computadores, servidores e software em um estado desejado e consistente.
6 - Automação de Testes	Uso de software para controlar a execução de testes, a comparação de resultados esperados com resultados reais, e a configuração de pré-condições de testes.

7- Versionamento Semântico	Convenção para nomear e gerenciar versões de software de forma a comunicar o impacto das mudanças no código.
8- Balanceamento de Carga	Distribuição automática de tráfego de rede ou pedidos entre vários servidores.
9- Código de Infraestrutura	Código que cria e configura a infraestrutura necessária para uma aplicação.
10- Dashboard de Monitoramento	Interface visual que exibe métricas importantes da aplicação e da infraestrutura.

Machine Learning Operations (MLOps)

Machine Learning Operations (MLOps) é uma área de atuação profissional responsável por dar suporte ao modelos, ao desenvolvimento e à operacionalização do ciclo de vida de Machine Learning estruturado nos princípios e práticas de DevOps.

Um fluxo de trabalho (workflow) de MLOps muito comum é composto por:

1. Extração de Dados (Data Extraction): etapa caracterizada pela integração de dados relevantes de fontes variadas.
2. Análise de Dados (Data Analysis): etapa caracterizada pela compreensão dos dados existentes nos conjuntos de dados.
3. Limpeza de Dados, Transformação e Engenharia de Atributos (Data Cleaning, Transformation and Feature Engineering): etapa caracterizada pela divisão e preparação dos conjuntos de dados de treinamento, validação e teste.
4. Treinamento do Modelo (Model Training): etapa caracterizada pelo treinamento de modelos de Machine Learning e armazenamento dos modelos com melhor desempenho, partindo de diferentes algoritmos e configurações de parâmetros.
5. Validação do Modelo (Model Validation): etapa caracterizada pela avaliação interativa da qualidade do modelo no conjunto de dados de teste e pela constatação de que se o modelo está atendendo aos critérios de qualidade baseada nas métricas de desempenho.
6. Serviço do Modelo (Model Serving): etapa caracterizada pela implantação dos modelos nos ambientes alvo integrados a outros componentes de software.

7. Monitoramento do Modelo (Model Monitoring): etapa caracterizada pela detecção da degradação do modelo através de análises de uso, dados de entrada e desempenho.

Termo	Definição
1- Open source/Código aberto	O código do software é público e disponível para uso, modificação e distribuição.
2- Escalabilidade	A capacidade de aumentar o tamanho da carga de trabalho dentro da infraestrutura existente (hardware, software, etc.) sem impactar o desempenho.
3- Elasticidade	A capacidade de expandir ou reduzir dinamicamente os recursos da infraestrutura (computacional) conforme necessário para se adaptar às mudanças na carga de trabalho de maneira autônoma.
4- Cloud agnostic	O desempenho é consistente, independentemente da plataforma em que é implantado.
5- Extensibilidade	Defina facilmente seus próprios operadores, executores e amplie a biblioteca para que ela se ajuste ao nível de abstração adequado ao seu ambiente.
6- Gestão/Coleta de metadados	A gestão de metadados é usada para coletar dados durante todo o pipeline de ML.
7- Isolamento/Fraco acoplamento	Os componentes podem ser desenvolvidos e implantados independentemente e devem depender uns dos outros na menor medida possível.
8- CI/CD	A plataforma suporta Integração Contínua (CI) e Entrega Contínua (CD) para o pipeline completo de ML.

9- UI	Interface de Usuário ou Dashboard.
10- CLI	Interface de Linha de Comando.
11- API gateway	Em vez de chamar os serviços diretamente, os clientes podem chamar o gateway de API, que encaminha a chamada para os serviços apropriados no back-end e serve como ponto de entrada para os clientes.
12- DAGs	Grafos Acíclicos Dirigidos são usados para descrever o fluxo de trabalho ou podem ser encapsulados dentro da plataforma.
13- Data streaming (real-time)	O fluxo contínuo de dados gerados por várias fontes de dados é suportado e pode ser processado, armazenado, analisado e utilizado diretamente.
14- Data storage	Um banco de dados integrado para armazenar dados brutos, projetos e metadados.
15- Data analysis	Um componente do pipeline gera estatísticas de características tanto para dados de treinamento quanto para dados de serviço, que podem ser usados por outros componentes do pipeline.
16- Data transformation	Um componente do pipeline identifica anomalias nos dados de treinamento e de serviço e prepara os dados para tarefas de ML. O resultado deste passo são as divisões de dados.
17- Data monitoring	Os dados são monitorados para manter a qualidade e inspecionar métricas gerais.
18- API endpoint	A saída da gestão de dados pode ser acessada usando um gateway de API, que encaminha os dados, metadados ou esquema de dados.
19- Automação	O processo de gestão de dados pode ser

	executado automaticamente em produção com base em uma programação ou em resposta a um gatilho.
20- Library agnostic	Todos os principais frameworks e bibliotecas de ML são suportados.
21- Model tracking	O desempenho do modelo de ML intermediário pode ser rastreado e registrado para manter a reprodutibilidade e obter insights.
22- Model registry	Um repositório centralizado usado para padronizar a definição, armazenamento e acesso de características para treinamento e serviço, que é acessível via uma API.
23- Hyper parameter tuning	Um motor de otimização é encapsulado para o ajuste de hiperparâmetros para treinar os modelos de ML de forma eficiente.
24- Teste A/B	Testes A/B podem ser usados para rastrear diferenças entre duas versões de modelos preditivos ou modelos podem ser executados em paralelo em diferentes pontos de extremidade.
25- Detecção de anomalias	Os outliers são automaticamente identificados para revelar padrões irregulares do modelo de ML.
26- Detecção de drift	Mudanças significativas nas distribuições de dados e no desempenho da previsão são automaticamente detectadas para prevenir obsolescência e diminuição da precisão.
27- Alerta de threshold	É possível configurar alertas quando a distribuição de previsões varia significativamente dos valores esperados.
28- Monitoramento de performance	O desempenho preditivo do modelo é monitorado para potencialmente invocar

	uma nova iteração no processo de ML.

Tabela 1. Definições dos principais termos em MLOps

Infrastructure as Code (IaC)

Infrastructure as Code (IaC) é uma prática da área de DevOps que consiste na criação de documentos em linguagem de codificação descritiva de alto nível para automatizar o provisionamento da infraestrutura de TI.

IaC elimina a necessidade de configuração manual de servidores, sistemas operacionais, conexões de bancos de dados, entre outras tarefas.

As ferramentas de IaC mais conhecidas no mercado são: Ansible, Terraform, Pulumi, Azure Resource Manager (ARM) e Google Cloud Deployment Manager.

Terraform é uma ferramenta popular de Infrastructure as Code (IaC) usada para provisionar infraestrutura em várias plataformas, como AWS, Azure, Google Cloud, entre outras. Ele utiliza uma linguagem de configuração chamada HashiCorp Configuration Language (HCL) para descrever a infraestrutura desejada de forma declarativa. Abaixo está um exemplo básico de um documento de configuração do Terraform para provisionar uma instância EC2 na AWS:

```
provider "aws" {
  region = "us-west-2"
}

resource "aws_instance" "example" {
  ami           = "ami-abcdefgh"
  instance_type = "t2.micro"
}

output "ip" {
  value = aws_instance.example.public_ip
}
```

Bloco Provider:

- provider "aws" { ... }: Este bloco indica que você está usando o provedor AWS. O Terraform tem provedores para várias plataformas e serviços.
- region = "us-west-2": Especifica a região da AWS onde os recursos serão provisionados.

Bloco Resource:

- resource "aws_instance" "example" { ... }: Este bloco define um recurso, que neste caso é uma instância EC2 na AWS.
- ami = "ami-abcdefgh": Especifica a Amazon Machine Image (AMI) que será usada para lançar a instância.
- instance_type = "t2.micro": Especifica o tipo de instância que será lançado.

Bloco Output:

- output "ip" { ... }: Este bloco define uma saída que será exibida após o Terraform aplicar a configuração.
- value = aws_instance.example.public_ip: Especifica que o IP público da instância EC2 será exibido como saída.

Quando este código é aplicado usando o comando terraform apply, o Terraform cria uma instância EC2 na AWS com as especificações fornecidas. A saída do comando mostrará o IP público da instância EC2 criada.

APÊNDICE 3

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 16 de nov. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Esta entrega consistiu na finalização da construção do embasamento teórico dos trabalhos que serão desenvolvidos, além da realização do planejamento de atividades a serem executadas até o fim da Residência.

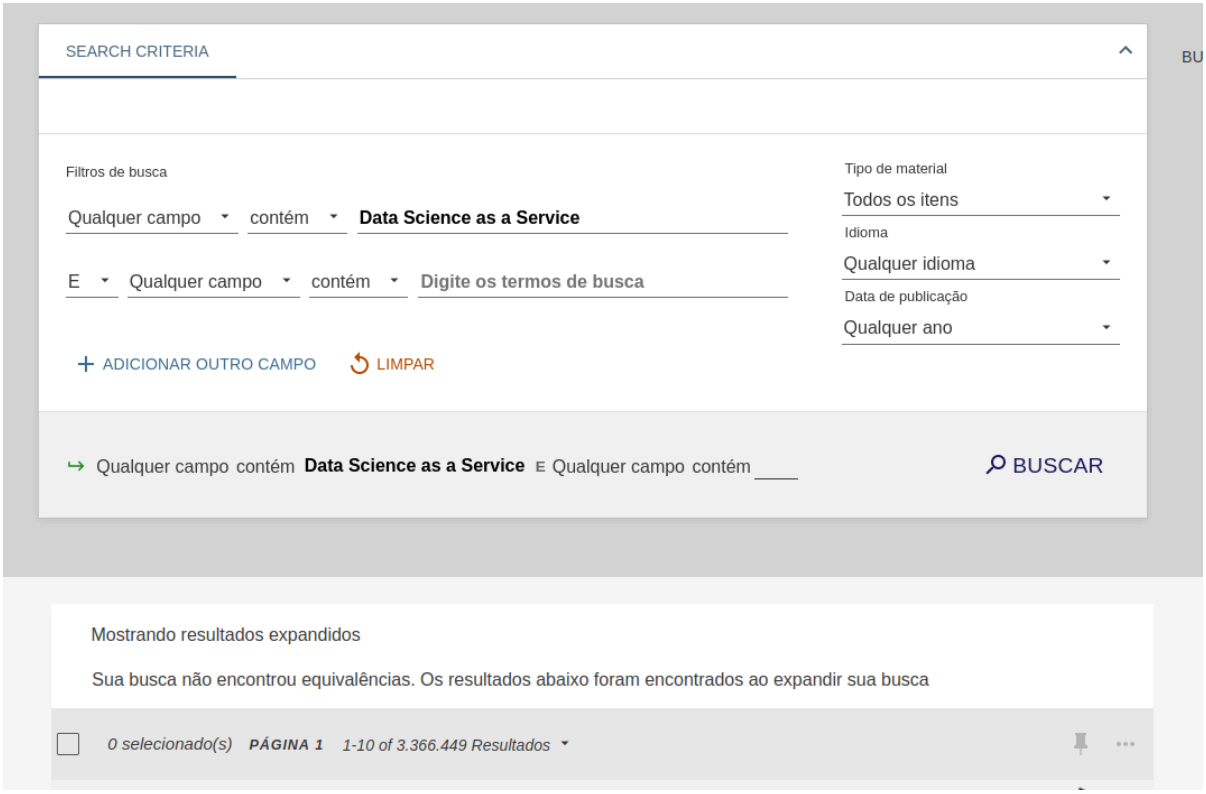
Os requisitos básicos para a entrega eram:

- Finalizar o vocabulário incluindo conceitos de DevOps.
- Buscar por repositórios de códigos úteis.
- Decidir sobre a aplicação de cada integrante e realizar o levantamento de requisitos.

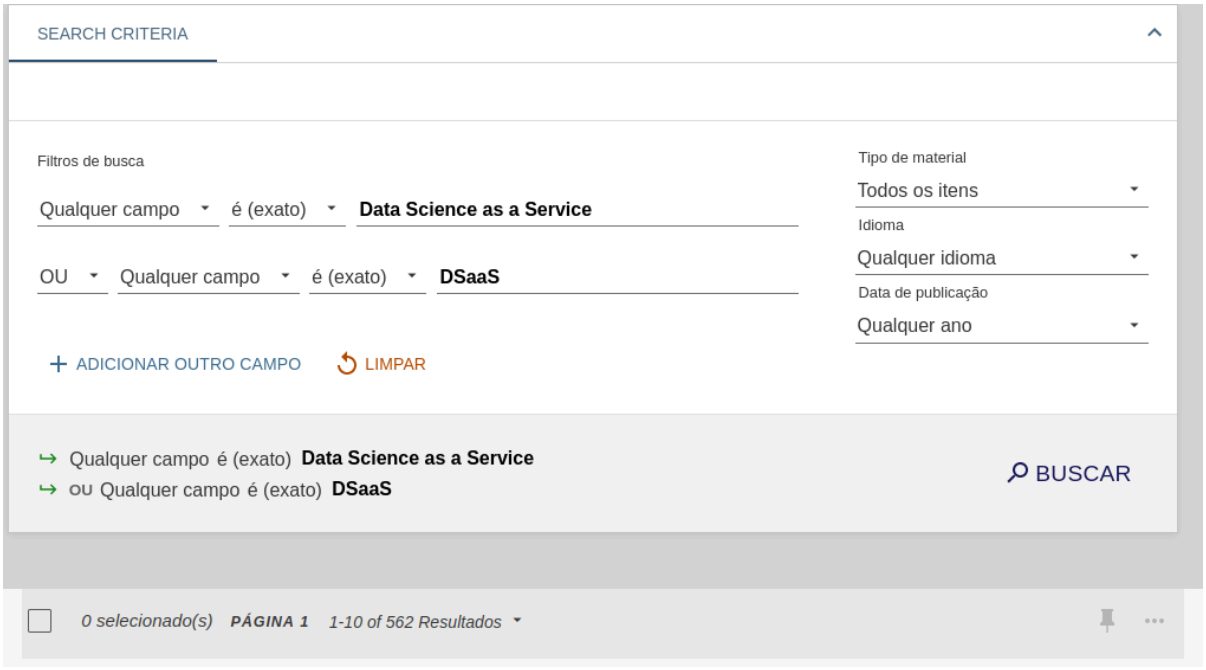
Os produtos gerados para esta entrega estão descritos a seguir:

- Como aplicação para residência foi decidido seguir com o tema de "Estudo sobre a construção de uma arquitetura de MLOps para o modelo de negócio Data Science as a Service (DSaaS)"
  - Abordagem "Top Down" - Listagem de modelos de negócio e ponderações que suscitaram a escolha do modelo DSaaS
    - Link
  - Pouca documentação:

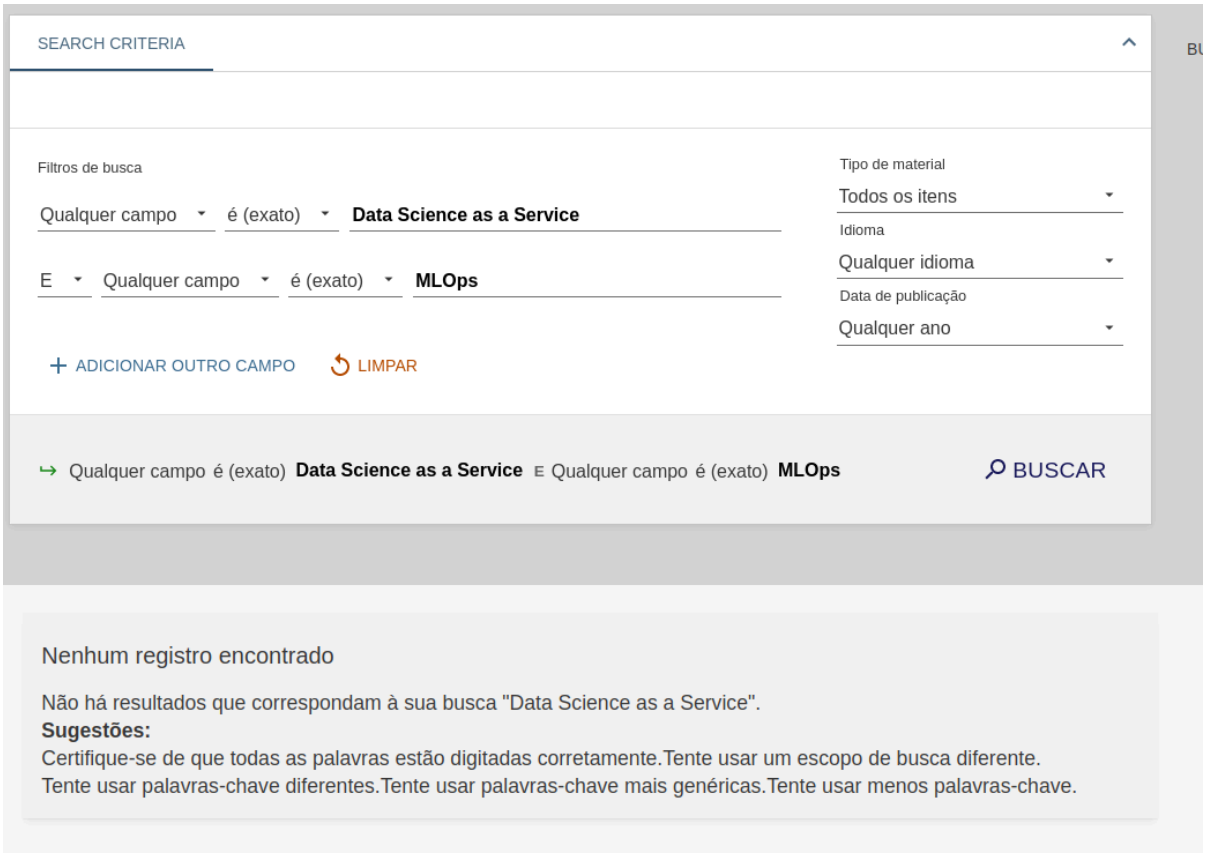
Primeira Pesquisa:



Segunda filtragem:



Terceira Filtragem:



- Estruturação do processo de revisão bibliográfica utilizando a ferramenta Parsifal.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

Para a próxima entrega do dia 23/11/2023, estão planejadas as seguintes atividades:

- Elaboração de cronograma para desenvolvimento de pesquisas
- Busca de repositórios úteis para criação do protótipo da arquitetura

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Gate parcialmente elaborado em conjunto com Állan e Gabriel da Mata

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Estudo sobre Modelos de Negócio

Análise de Modelos de Negócio "As a Service" e a Adequação da Arquitetura de MLOps

1. Software as a Service (SaaS)

Descrição: SaaS é um modelo de distribuição de software onde os aplicativos são hospedados por um provedor de serviços e disponibilizados aos clientes pela internet. É conhecido por sua conveniência e escalabilidade.

Ponderações Negativas para MLOps:

- **Foco em Software Genérico:** SaaS geralmente oferece soluções de software genéricas que podem não se beneficiar diretamente das complexidades e especificidades do MLOps.
- **Menor Necessidade de Personalização:** SaaS tende a ser menos personalizado em comparação com soluções baseadas em dados, limitando o impacto potencial do MLOps.

2. Platform as a Service (PaaS)

Descrição: PaaS fornece um ambiente de plataforma e ferramentas para permitir que os desenvolvedores criem e gerenciem aplicações web e móveis sem a complexidade de construir e manter a infraestrutura normalmente associada ao processo.

Ponderações Negativas para MLOps:

- **Amplitude vs. Profundidade:** PaaS fornece uma plataforma ampla, o que pode diluir o foco específico necessário para a implementação eficaz de MLOps.
- **Complexidade de Integração:** A integração de MLOps em uma plataforma existente pode ser complexa e custosa.

3. Infrastructure as a Service (IaaS)

Descrição: IaaS oferece recursos de computação, como servidores virtuais e armazenamento, através da internet. Com IaaS, os usuários podem alugar infraestrutura de TI em vez de comprar e instalá-la fisicamente.

Ponderações Negativas para MLOps:

- **Foco em Infraestrutura:** IaaS se concentra mais na infraestrutura de hardware e rede, que é apenas uma parte do quebra-cabeça do MLOps.
- **Desafios de Escalabilidade:** A escalabilidade de soluções de MLOps em IaaS pode ser desafiadora, exigindo investimentos significativos.

4. Communication as a Service (CaaS)

Descrição: CaaS é um modelo de serviço que permite aos usuários ter acesso a soluções de comunicação avançadas, como VoIP, mensagens instantâneas e videoconferências, sem a necessidade de investir em infraestrutura de comunicação própria.

Ponderações Negativas para MLOps:

- **Foco em Comunicação:** CaaS se concentra em soluções de comunicação, que têm pouca ou nenhuma relação com as necessidades de modelagem e operacionalização de machine learning.
- **Limitações de Aplicabilidade:** A aplicabilidade do MLOps em um contexto de CaaS é limitada e não explora plenamente suas capacidades.

5. Data Science as a Service (DSaaS)

Descrição: DSaaS envolve a oferta de análises de dados e insights de ciência de dados como um serviço. Empresas podem utilizar DSaaS para obter insights de dados sem a necessidade de manter uma equipe interna de cientistas de dados.

Ponderações Positivas para MLOps:

- **Alinhamento com Ciência de Dados:** DSaaS está intrinsecamente ligado à ciência de dados, tornando-o um candidato ideal para a implementação de MLOps.
- **Necessidade de Ciclo Contínuo de ML:** DSaaS se beneficia diretamente do ciclo contínuo de desenvolvimento, teste e implantação que o MLOps facilita.
- **Produtização de Modelos de Dados:** MLOps oferece o caminho para transformar modelos de dados em produtos viáveis, alinhando-se com a natureza de negócio do DSaaS.

- **Inovação e Competitividade:** A adoção de MLOps em DSaaS coloca uma organização na vanguarda da inovação, oferecendo serviços de ciência de dados mais eficientes e competitivos.
- **Pouca Documentação Existente:** A escassez de literatura sobre a combinação de DSaaS com MLOps apresenta uma oportunidade única de exploração e contribuição acadêmica.

APÊNDICE 4

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 23 de nov. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Esta entrega consistiu na finalização da construção do embasamento teórico dos trabalhos que serão desenvolvidos, além da realização do planejamento de atividades a serem executadas até o fim da Residência.

Os produtos gerados para esta entrega:

- Construção de documento "O que é Data Science as a Service?"
- Cronograma de desenvolvimento de atividades
- Procura de modelos de negócio similares (Ou complementares)

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Definição de case prático para ser desenvolvido ao longo da residência
  - Prospecção de empresa

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Atividades parcialmente elaboradas com o Grupo de MLOps

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

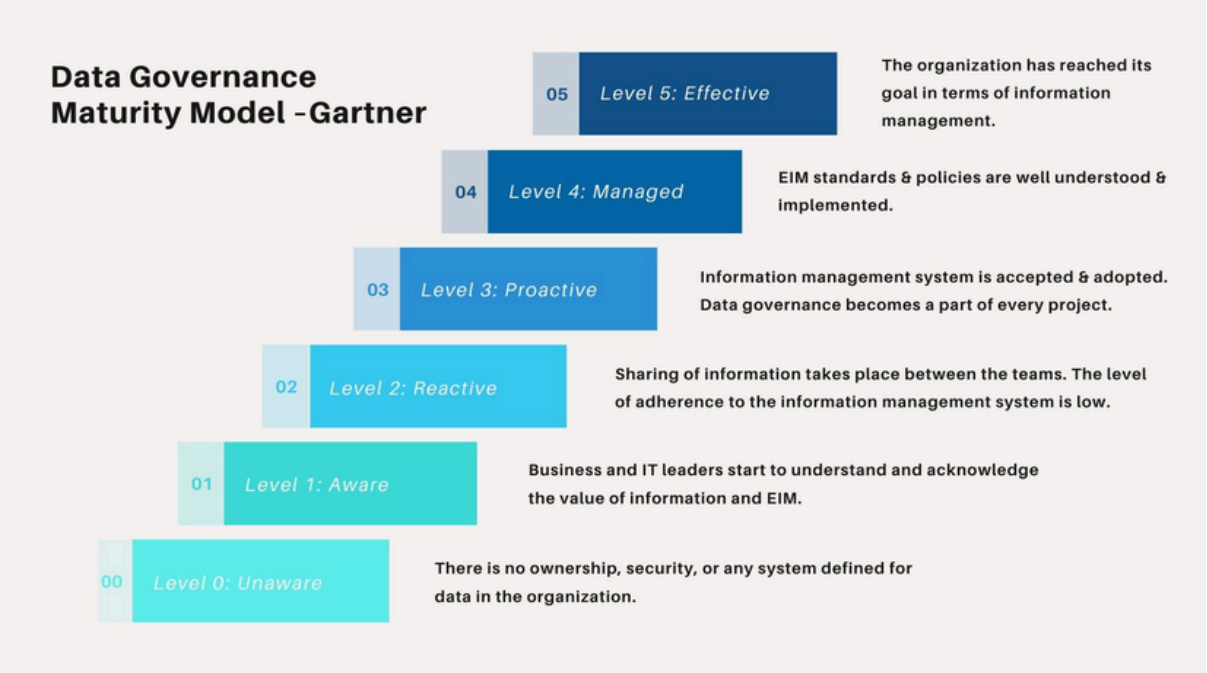
Estudo sobre Data Science as a Service (DSaaS)

Data Science as a Service (DSaaS): Um Modelo de Negócio Emergente

1. O que é Data Science as a Service (DSaaS)?

Data Science as a Service (DSaaS) é um modelo de negócio inovador que oferece serviços de ciência de dados através de uma abordagem "como serviço". Esse modelo permite que as empresas acessem insights avançados de dados e capacidades analíticas sem a necessidade de desenvolver internamente equipes e infraestruturas especializadas em ciência de dados. DSaaS abrange desde a coleta e processamento de dados até a análise avançada e modelagem preditiva, oferecendo soluções personalizadas conforme as necessidades específicas de cada cliente.

2. Aplicabilidade do DSaaS Segundo o Modelo Gartner da IBM de Maturidade em Dados



O Modelo Gartner da IBM descreve diferentes níveis de maturidade de dados nas empresas, e DSaaS se encaixa particularmente bem nos seguintes níveis:

Nível 2 / Reativo

- **Descrição:** Neste nível, as empresas reconhecem a importância dos dados, mas estão apenas começando a utilizá-los de forma eficaz. A coleta de dados ocorre, mas a transformação desses dados em insights acionáveis é um desafio. A estratégia de dados e os recursos dedicados à análise são limitados ou inexistentes.
- **Aplicabilidade do DSaaS:** DSaaS pode fornecer a essas empresas o suporte necessário para começar a explorar dados de maneira mais estratégica, sem a necessidade de um grande investimento inicial em recursos internos.

Perguntas para Autoavaliação:

- A sua empresa coleta dados, mas enfrenta dificuldades para utilizá-los estrategicamente?
- Existe uma falta de direção clara ou de recursos dedicados à análise de dados?
- Você sente que os dados são subutilizados no processo de tomada de decisão?

Nível 3 / Proativo

- **Descrição:** Empresas no nível proativo já implementaram sistemas de gestão da informação e estão em processo de aprimorar a utilização de dados. Há um esforço para desenvolver uma cultura de dados, mas ainda existem oportunidades significativas para melhorar a análise e a utilização dos dados.
- **Aplicabilidade do DSaaS:** DSaaS pode acelerar o processo de adoção e aprimoramento de análises de dados, fornecendo expertise e tecnologias avançadas que podem não estar disponíveis internamente.

Perguntas para Autoavaliação:

- Sua empresa já possui sistemas de gestão de informação implementados?
- Você está buscando formas de aprimorar a análise de dados para suportar decisões de negócios?
- Existe uma cultura de dados emergente, mas com espaço para melhorias na análise e utilização efetiva dos dados?

Nível 4 / Gerenciado

- **Descrição:** Neste estágio, as empresas têm padrões e políticas de dados bem estabelecidos. Elas buscam otimizar e expandir suas capacidades de análise de dados. A integração de dados é eficaz em todas as operações de negócios, com um foco contínuo em inovação e melhoria contínua.
- **Aplicabilidade do DSaaS:** DSaaS oferece a essas organizações a oportunidade de escalar suas operações de dados e incorporar novas técnicas e modelos analíticos, mantendo o foco em suas competências principais.

Perguntas para Autoavaliação:

- Sua empresa possui padrões e políticas de dados estabelecidos e busca constantemente otimizá-los?
- Você está procurando expandir suas capacidades de análise de dados?
- Existe uma integração efetiva de dados em todas as operações de negócios, com um foco em inovação e melhoria contínua?

3. Análise de Mercado para DSaaS

Tamanho do Mercado e Taxa de Crescimento

- **Tamanho do Mercado em 2021**: O mercado global de plataformas de ciência de dados foi avaliado em USD 95,3 bilhões.
- **Previsão para 2026**: Espera-se que o mercado atinja USD 322,9 bilhões.
- **Taxa de Crescimento Anual Composta (CAGR)**: O mercado está projetado para crescer a uma CAGR de 27,7% de 2021 a 2026.

Fatores Impulsionadores do Mercado

- **Crescimento de Big Data**: O aumento significativo no volume de dados gerados, especialmente dados não estruturados, está impulsionando a demanda por plataformas de ciência de dados.
- **Adoção de Soluções Baseadas em Nuvem**: Há um aumento na adoção de soluções baseadas em nuvem, o que está contribuindo para o crescimento do mercado.
- **Necessidade de Insights Profundos**: As empresas estão buscando extrair insights mais profundos de grandes volumes de dados para obter vantagens competitivas.

Desafios do Mercado

- **Falta de Força de Trabalho Qualificada**: Há uma carência de profissionais qualificados em ciência de dados, o que pode ser um desafio para o crescimento do mercado.

Conclusão

Com base nas informações acima, o mercado de DSaaS representa uma oportunidade de negócio significativa e está em uma trajetória de crescimento rápido. As empresas em diferentes níveis de maturidade de dados podem se beneficiar de maneiras variadas ao adotar soluções de DSaaS.

Fontes

- Markets and Markets - Data Science Platform Market
- https://ilumeo.com.br/categorias/2022-01-05-medindo-o-nivel-de-maturidade-dos-dados-em-sua-empresa/

APÊNDICE 5

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 30 de nov. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Os requisitos básicos para a entrega eram:

- Prospecção de empresa parceira para desenvolvimento de projeto
- Definição de case prático para ser desenvolvido

Os produtos gerados para esta entrega estão descritos a seguir:

- Parceria com a empresa Data H para realização de estudo sobre modelo de negócio.
- Participação e implementação em um projeto prático da empresa.

- Além disso, durante esse estudo de mercado foi percebido que a pesquisa realizada teria maior valor agregada na etapa de integração entre o ciclo de ciência de dados e a etapa de deployment, como representado a seguir:

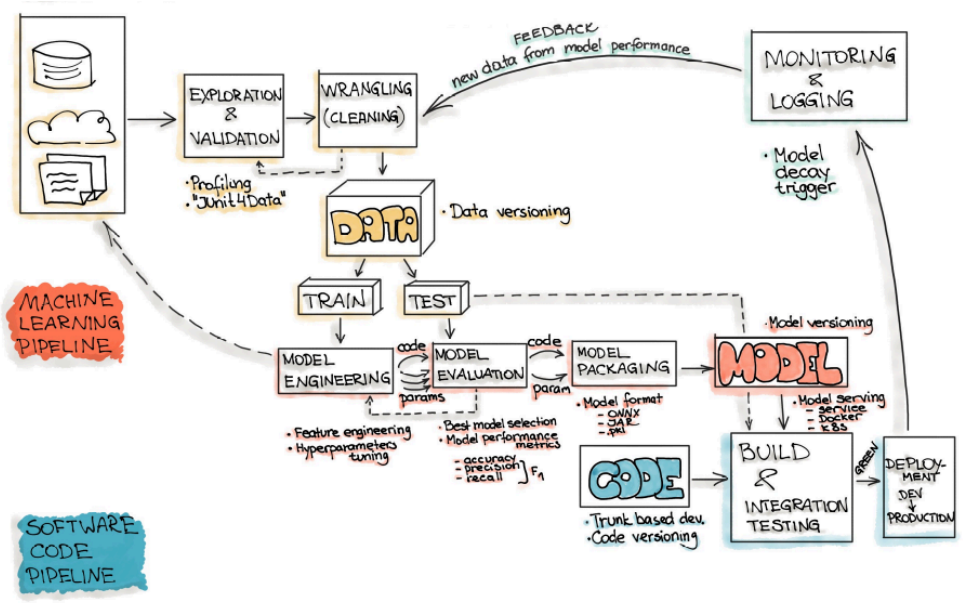


figura obtida do canal data professor

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

Para a próxima entrega do dia 07/12/2023, estão planejadas as seguintes atividades:

- Elaboração do esboço da arquitetura em cloud.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Este Gate foi desenvolvido tendo cooperação dos colaboradores da Data H:
- Evandro Barros, CEO
- Celso Azevedo, Co-Fundador
- Marcelo Piovan, CTO

Neste gate, o Professor Aldo André Díaz Salazar esteve na banca avaliadora substituindo a Professora Luana.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Em análise!

APÊNDICE 6

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 7 de dez. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Os requisitos básicos para a entrega eram:

- Elaboração do esboço da arquitetura em Cloud

Os produtos gerados para esta entrega estão descritos a seguir:

- Estudo de ferramentas disponíveis para construção de arquitetura em Cloud
- Construção do esboço da arquitetura listando ferramentas
  - LINK
- Análise de ferramentas necessárias para cada etapa da construção

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega do dia 14/12/2023, estão planejadas as seguintes atividades:

- início da construção da arquitetura em cloud.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Neste gate, o Professor Aldo André Díaz Salazar esteve na banca avaliadora substituindo a Professora Luana.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Em análise!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

Esboço de arquitetura de MLOps desejada

Focando especificamente nos momentos de conexão após o ciclo CRISP-DM e na produtização, podemos considerar os seguintes componentes e fluxos:

1. Conexão Pós-Ciclo CRISP-DM

- **Repositório de Dados:** Uma infraestrutura de armazenamento de dados escalável e segura na nuvem para armazenar os conjuntos de dados processados e refinados após as etapas do CRISP-DM.
- **Ambiente de Modelagem:** Plataformas de modelagem e análise de dados na nuvem, onde os modelos de machine learning são desenvolvidos, treinados e validados.
- **Integração de Dados:** Ferramentas para integrar dados de diversas fontes e formatos, facilitando a preparação e transformação de dados para modelagem.

2. Produtização

- **Serviços de Implantação de Modelos:** Plataformas na nuvem que permitem a implantação rápida e eficiente de modelos de machine learning, com suporte para escalabilidade e gerenciamento de carga.
- **APIs e Interfaces de Acesso:** Criação de APIs para permitir que os modelos sejam acessados e utilizados por aplicações de clientes ou outros sistemas.
- **Monitoramento e Manutenção:** Ferramentas para monitorar o desempenho dos modelos em produção, identificar problemas e realizar manutenções e atualizações necessárias.

3. Componentes Adicionais

- **Segurança e Conformidade:** Implementação de protocolos de segurança robustos para proteger dados e modelos, além de garantir conformidade com regulamentações de dados.
- **Orquestração e Automação:** Utilização de ferramentas de orquestração para automatizar fluxos de trabalho de data science e MLOps, melhorando a eficiência e reduzindo erros manuais.
- **Escalabilidade e Flexibilidade:** Capacidade de escalar recursos conforme a demanda, garantindo flexibilidade e otimização de custos.

4. Fluxo de Trabalho Sugerido para Arquitetura em Cloud de MLOps com DSaaS

1. Coleta e Preparação de Dados

- **Ferramentas:** AWS Glue, Google Cloud DataPrep, Azure Data Factory.
- **Justificativa:** Estas ferramentas oferecem capacidades robustas de ETL (Extract, Transform, Load), permitindo a coleta, limpeza e transformação de dados de diversas fontes. São escaláveis e integram-se bem com outras soluções em cloud.

2. Desenvolvimento de Modelos

- **Ferramentas:** Jupyter Notebooks em AWS SageMaker, Google AI Platform Notebooks, Azure Machine Learning.
- **Justificativa:** Estes ambientes fornecem plataformas interativas para desenvolvimento e treinamento de modelos de machine learning, com acesso a recursos computacionais escaláveis e integração com armazenamento de dados.

3. Validação e Testes

- **Ferramentas:** TensorFlow Extended (TFX), MLflow.
- **Justificativa:** TFX e MLflow oferecem componentes para validação de modelos, incluindo testes de qualidade de dados e desempenho do modelo. Eles ajudam a garantir que os modelos sejam confiáveis e prontos para produção.

4. Implantação

- **Ferramentas:** Kubernetes, Docker, AWS Elastic Container Service (ECS), Azure Kubernetes Service (AKS).
- **Justificativa:** Estas ferramentas facilitam a implantação de modelos em ambientes de produção, oferecendo gerenciamento de contêineres, escalabilidade e balanceamento de carga.

5. Acesso e Utilização

- **Ferramentas:** APIs RESTful, AWS API Gateway, Azure API Management.
- **Justificativa:** A criação de APIs RESTful permite que os modelos sejam facilmente acessados e integrados em aplicações de clientes. Ferramentas como AWS API Gateway e Azure API Management simplificam a criação, monitoramento e segurança das APIs.

6. Monitoramento e Otimização

- **Ferramentas:** Prometheus, Grafana, Amazon CloudWatch, Azure Monitor.

- **Justificativa:** Estas ferramentas oferecem monitoramento em tempo real do desempenho dos modelos e da infraestrutura subjacente. Elas permitem a detecção rápida de problemas e a otimização contínua do desempenho.

Considerações Adicionais

- **Segurança e Conformidade:** É crucial garantir a segurança dos dados e modelos, utilizando ferramentas como AWS Identity and Access Management (IAM) ou Azure Active Directory para gerenciar acessos e permissões.
- **Documentação e Versionamento:** Ferramentas como Git e DVC (Data Version Control) são importantes para manter a documentação e o versionamento dos modelos e dados.

APÊNDICE 7

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 14 de dez. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Os requisitos para essa entrega era realizar um estudo detalhado sobre as plataformas de cloud computing Amazon Web Services (AWS) e Google Cloud Platform (GCP)

Produtos Gerados

- Foi entregue um documento detalhado que descreve os seguintes aspectos de cada plataforma:
  - Link
- Documento explicando como está sendo realizado planejamento da implementação de uma arquitetura de MLOPS para DSaaS com os frameworks de cada plataforma:
  - Link

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega, prevista para o dia 21/12, estão planejadas as seguintes atividades:

- **Avaliação Comparativa entre AWS e GCP:**
  - Realizar uma análise comparativa entre as soluções de DSaaS propostas com AWS e GCP, destacando vantagens, desvantagens e casos de uso específicos para cada plataforma.
  - Desenvolver critérios de avaliação baseados em custo, desempenho, facilidade de uso e integração.
- **Prototipagem:**
  - Continuar a prototipagem de uma solução de DSaaS usando a plataforma selecionada (AWS ou GCP) com base na análise comparativa.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Comparativo entre AWS e Google Cloud para MLOps e DSaaS

Amazon Web Services (AWS)

Visão Geral

Amazon Web Services é uma das plataformas de cloud computing mais abrangentes e amplamente adotadas no mundo. Oferece uma vasta gama de serviços integrados que vão desde computação em nuvem, armazenamento, bancos de dados até machine learning e analytics.

Principais Serviços para MLOps e DSaaS

- **Amazon S3:** Serviço de armazenamento altamente escalável, ideal para armazenar grandes volumes de dados.
- **Amazon EC2:** Oferece capacidade computacional escalável na nuvem, crucial para treinamento e execução de modelos de machine learning.
- **AWS Lambda:** Permite a execução de código em resposta a eventos, útil para automação de tarefas em MLOps.
- **Amazon SageMaker:** Uma plataforma completa para criar, treinar e implantar modelos de machine learning, facilitando todo o ciclo de vida do ML.

Amazon Web Services (AWS)

AWS S3

- **Framework:** Armazenamento de objetos com alta durabilidade e disponibilidade. Suporta armazenamento de grandes volumes de dados, como datasets de machine learning.
- **Funcionalidades:** Oferece recursos como versionamento, lifecycle policies e integração com AWS Lambda para processamento de dados.

AWS EC2

- **Framework:** Serviço de computação em nuvem que permite executar aplicações em servidores virtuais.

- **Funcionalidades:** Oferece uma ampla gama de tipos de instâncias, incluindo instâncias otimizadas para computação, memória ou I/O, e suporte para GPUs para treinamento de modelos de ML.

AWS Lambda

- **Framework:** Serviço de computação sem servidor que executa código em resposta a eventos.
- **Funcionalidades:** Ideal para automação de tarefas em MLOps, como triggers para processos de CI/CD ou pipelines de dados.

Amazon SageMaker

- **Framework:** Plataforma completa para machine learning que permite aos cientistas de dados e desenvolvedores criar, treinar e implantar modelos de ML rapidamente.
- **Funcionalidades:** Inclui Jupyter Notebooks integrados, treinamento e tuning de modelos, e um ambiente para implantação fácil de modelos.

Por que AWS para DSaaS?

- **Escala e Flexibilidade:** AWS oferece uma infraestrutura global que pode escalar conforme a demanda do projeto, essencial para DSaaS que requer alta disponibilidade e desempenho.
- **Integração e Automação:** A vasta gama de serviços AWS permite a integração e automação eficiente de processos de MLOps, desde a coleta de dados até a implantação de modelos.
- **Segurança e Confiabilidade:** AWS é conhecida por sua segurança robusta e conformidade com padrões globais, garantindo a proteção de dados e modelos.

Google Cloud Platform (GCP)

Visão Geral

Google Cloud Platform é uma suíte de cloud computing que oferece serviços de hospedagem na infraestrutura do Google. É conhecida por suas capacidades avançadas em machine learning, análise de dados e escalabilidade.

Principais Serviços para MLOps e DSaaS

- **Google Cloud Storage:** Para armazenamento de dados escalável e seguro.

- **Google Compute Engine:** Oferece máquinas virtuais personalizáveis para cargas de trabalho de computação intensiva.
- **Google Kubernetes Engine (GKE):** Ideal para gerenciamento de aplicativos em contêineres, facilitando a implantação e escalabilidade.
- **Google AI Platform:** Fornece um ambiente integrado para o desenvolvimento de modelos de machine learning, desde a preparação de dados até a implantação.

Google Cloud Platform (GCP)

Google Cloud Storage

- **Framework:** Armazenamento de objetos altamente escalável e seguro.
- **Funcionalidades:** Suporta armazenamento de dados estruturados e não estruturados, com forte consistência de dados e integração com outros serviços do GCP.

Google Compute Engine

- **Framework:** Serviço de VMs que oferece máquinas virtuais personalizáveis.
- **Funcionalidades:** Permite configurações customizadas de CPU, memória, disco e GPUs, adequadas para diferentes cargas de trabalho de ML.

Google Kubernetes Engine (GKE)

- **Framework:** Serviço gerenciado para implantação, gerenciamento e escalonamento de aplicações em contêineres.
- **Funcionalidades:** Integração com o ecossistema Kubernetes, auto-scaling e balanceamento de carga para aplicações distribuídas.

Google AI Platform

- **Framework:** Plataforma integrada para o desenvolvimento de modelos de ML, desde a preparação de dados até a implantação.
- **Funcionalidades:** Inclui serviços para treinamento e previsão de modelos, suporte para TensorFlow, PyTorch, e outras frameworks de ML, e ferramentas para o gerenciamento do ciclo de vida do modelo.

Por que GCP para DSaaS?

- **Inovação em AI e ML:** GCP é reconhecida por suas inovações em AI e ML, oferecendo ferramentas avançadas e pré-treinadas que podem acelerar o desenvolvimento em DSaaS.
- **Análise de Dados em Grande Escala:** Com ferramentas como BigQuery, GCP permite a análise de grandes volumes de dados, essencial para insights em DSaaS.
- **Rede Global e Desempenho:** A infraestrutura global do Google garante alta disponibilidade e desempenho, crucial para serviços de DSaaS que exigem resposta rápida e confiável.

Implementação de DSaaS com AWS

Fase 1: Coleta e Armazenamento de Dados

- **AWS S3:** Utilizado para armazenar grandes volumes de dados brutos e processados.
- **AWS Glue:** Para processos ETL, transformando e preparando dados para análise.

Fase 2: Desenvolvimento e Treinamento de Modelos

- **Amazon SageMaker:** Ambiente central para desenvolvimento, treinamento e tuning de modelos de machine learning.
- **AWS EC2:** Instâncias com GPUs para treinamento intensivo de modelos de ML.

Fase 3: Implantação e Monitoramento

- **AWS Lambda e AWS Step Functions:** Para automação de workflows de MLOps, como a implantação automatizada de modelos.
- **Amazon CloudWatch:** Monitoramento do desempenho dos modelos e da infraestrutura.

Fase 4: Acesso e Integração

- **AWS API Gateway:** Criação de APIs RESTful para permitir que aplicações de clientes acessem os modelos de ML.

Implementação de DSaaS com GCP

Fase 1: Coleta e Armazenamento de Dados

- **Google Cloud Storage:** Armazenamento escalável para dados brutos e processados.
- **Google Cloud Dataflow:** Para processamento e transformação de dados em larga escala.

Fase 2: Desenvolvimento e Treinamento de Modelos

- **Google AI Platform:** Plataforma integrada para desenvolvimento, treinamento e avaliação de modelos de ML.
- **Google Compute Engine:** Uso de VMs personalizadas com GPUs para treinamento de modelos.

Fase 3: Implantação e Monitoramento

- **Google Kubernetes Engine (GKE):** Para gerenciar e escalar modelos de ML em contêineres.
- **Google Cloud Monitoring e Logging:** Para monitorar o desempenho dos modelos e da infraestrutura.

Fase 4: Acesso e Integração

- **Google Endpoints e Apigee:** Para criar e gerenciar APIs RESTful que facilitam o acesso aos modelos de ML por aplicações de clientes.

APÊNDICE 8

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 21 de dez. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Requisitos da Entrega

Os requisitos para esta entrega envolviam a realização de um estudo detalhado sobre as plataformas de cloud computing Amazon Web Services (AWS) e Google Cloud Platform (GCP). O foco era entender como cada uma pode ser utilizada para implementar uma solução de Data Science as a Service (DSaaS) com ênfase em MLOps.

Produtos Gerados

Foram entregues documentos detalhados que descrevem os seguintes aspectos de cada plataforma:

Amazon Web Services (AWS):

- **Visão Geral dos Serviços Relevantes para DSaaS e MLOps:** Uma análise abrangente dos serviços AWS, incluindo AWS S3, EC2, Lambda e SageMaker.
- **Esboço de Implementação para DSaaS:** Detalhamento de como os serviços AWS podem ser integrados para criar uma solução eficaz de DSaaS, com ênfase em MLOps.

Google Cloud Platform (GCP):

- **Descrição Detalhada dos Serviços para DSaaS e MLOps:** Uma exploração profunda dos serviços GCP, como Google Cloud Storage, Compute Engine, Kubernetes Engine e AI Platform.
- **Esboço de Implementação para DSaaS:** Como a GCP pode ser utilizada para desenvolver uma solução de DSaaS, considerando as particularidades da plataforma.

Link para o Documento: LINK

Esse documento oferece insights sobre as capacidades de cada plataforma e como elas podem ser

aplicadas especificamente para o desenvolvimento de soluções de DSaaS em um ambiente de MLOps.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega, estão planejadas as seguintes atividades:

- Continuar a prototipagem de uma solução de DSaaS usando as plataformas selecionadas (AWS ou GCP) com base na análise comparativa.
- Documentar o processo de configuração inicial, desafios encontrados e soluções adotadas.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Desculpe o atraso na entrega, tive um imprevisto familiar e com isso acabei deixando passar o horário de entrega.

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Análise de Frameworks em AWS e GCP para DSaaS e MLOps

Amazon Web Services (AWS)

AWS S3

- **Feedback:** Ao testar o AWS S3, fiquei impressionado com sua escalabilidade e durabilidade. A facilidade de integração com outros serviços AWS, como AWS Lambda, torna-o ideal para armazenamento de dados em DSaaS. No entanto, a estrutura de preços pode ser complexa, especialmente ao lidar com grandes volumes de dados.

AWS EC2

- **Feedback:** A flexibilidade das instâncias EC2 é notável. A capacidade de escolher entre diferentes tipos de instâncias, incluindo aquelas otimizadas para ML, oferece grande versatilidade. Contudo, gerenciar o custo e a escalabilidade pode ser desafiador para usuários menos experientes.

AWS Lambda

- **Feedback:** A utilização do AWS Lambda revelou-se extremamente eficiente para automação de tarefas e execução de código em resposta a eventos. Sua natureza sem servidor e modelo de cobrança baseado em uso são pontos fortes. No entanto, há limitações em termos de tempo de execução e memória que precisam ser consideradas.

Amazon SageMaker

- **Feedback:** SageMaker provou ser uma ferramenta poderosa para todo o ciclo de vida do ML. Sua integração com Jupyter Notebooks e facilidade de implantação de modelos são impressionantes. Porém, a curva de aprendizado pode ser íngreme para iniciantes em ML.

Google Cloud Platform (GCP)

Google Cloud Storage

- **Feedback:** Durante os testes, o Google Cloud Storage mostrou-se altamente confiável e escalável. A integração com outros serviços do GCP é um ponto forte. No entanto, assim como o AWS S3, a estrutura de preços pode ser complexa.

Google Compute Engine

- **Feedback:** A personalização oferecida pelo Google Compute Engine é excelente, especialmente para cargas de trabalho específicas de ML. A possibilidade de usar GPUs é um grande benefício. A interface de usuário, no entanto, pode ser um pouco menos intuitiva do que a da AWS.

Google Kubernetes Engine (GKE)

- **Feedback:** GKE se destacou na gestão de aplicações em contêineres. Sua integração com o ecossistema Kubernetes e recursos de auto-scaling são notáveis. A complexidade de configuração, porém, pode ser um obstáculo para novos usuários.

Google AI Platform

- **Feedback:** A AI Platform é uma solução abrangente para o desenvolvimento de modelos de ML. A integração com TensorFlow e outras ferramentas de ML é um grande ponto positivo. Contudo, assim como o SageMaker, há uma curva de aprendizado associada.

Conclusão

Minha análise prática dos frameworks da AWS e GCP revelou que ambas as plataformas oferecem um conjunto robusto de ferramentas para DSaaS e MLOps. A escolha entre elas deve considerar fatores como familiaridade com a plataforma, requisitos específicos do projeto e considerações de custo. A AWS oferece uma gama mais ampla de serviços, enquanto a GCP se destaca em inovações específicas de AI e ML.

APÊNDICE 9

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 11 de jan. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Requisitos da Entrega

Os requisitos para esta entrega envolviam trazer os resultados obtidos ao longo das 10 semanas de pesquisa e a finalização da prototipação

Produtos Gerados

Portanto, foi gerado um documento que analisa os aspectos de importância para o modelo de negócio DSAAS, juntamente com uma estrutura sobre aspectos importantes de MLOPS para integração no modelo de negócio.

Produto gerado:

- Considerações finais DSAAS

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Organizar e adequar as entregas para o formato do TCC.
- Montar um repositório no Github com análises

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

Considerações Finais

Considerações Finais: Data Science as a Service e MLOps

Introdução

O campo da Ciência de Dados tem evoluído rapidamente, levando ao surgimento de modelos de negócios inovadores, como o Data Science as a Service (DSaaS). Este modelo oferece serviços de análise de dados sob demanda, permitindo que empresas de diferentes tamanhos e setores aproveitem os insights orientados por dados sem a necessidade de desenvolver internamente capacidades avançadas de análise. Em paralelo, a prática de MLOps (Machine Learning Operations) tem se destacado, focando na eficiência e automação do ciclo de vida dos modelos de Machine Learning.

Rentabilidade do DSaaS

O DSaaS emergiu como um modelo de negócio altamente rentável devido à sua flexibilidade, escalabilidade e capacidade de fornecer insights valiosos com custo-efetividade. Empresas podem acessar soluções avançadas de análise de dados sem os custos associados à construção e manutenção de uma infraestrutura de dados interna. Essa abordagem não só reduz custos operacionais, mas também acelera a tomada de decisão baseada em dados, um diferencial competitivo no mercado atual.

Emergência do DSaaS

A emergência do DSaaS como modelo de negócio se dá em um contexto onde os dados são considerados um ativo estratégico crucial. Com a crescente complexidade dos dados e a necessidade de análises mais profundas, o DSaaS oferece uma solução ágil e eficiente para as empresas se adaptarem rapidamente às mudanças do mercado e às novas demandas dos clientes. Além disso, a adoção crescente de tecnologias de cloud computing facilita a implementação deste modelo, tornando-o acessível a um espectro mais amplo de empresas.

Integração com MLOps

A integração do DSaaS com MLOps representa um avanço significativo na forma como os projetos de Machine Learning são gerenciados e implementados. O MLOps traz uma abordagem sistemática e automatizada para o ciclo de vida dos modelos de Machine Learning, garantindo maior eficiência, escalabilidade e qualidade. Essa integração permite que as soluções de DSaaS sejam não apenas mais robustas, mas também mais adaptáveis e alinhadas com as necessidades em constante evolução das empresas.

Análise das Soluções da AWS e Google Cloud para Diferentes Estágios de Desenvolvimento Empresarial

AWS e Empresas em Estágio Inicial

Para startups e empresas em estágio inicial, a AWS oferece vantagens como custo-efetividade e uma ampla gama de serviços. Suas soluções de armazenamento e computação em nuvem são ideais para empresas que precisam de escalabilidade rápida. Além disso, a AWS fornece uma ampla gama de ferramentas de aprendizado de máquina que são acessíveis até mesmo para equipes com conhecimento técnico limitado.

Google Cloud e Empresas em Expansão

A Google Cloud é particularmente vantajosa para empresas em fase de expansão. Seu forte investimento em IA e machine learning, combinado com a integração nativa com outras ferramentas do Google, como o BigQuery, oferece recursos avançados de análise de dados. Essas características são benéficas para empresas que estão expandindo suas operações e necessitam de insights mais profundos e análises preditivas.

Conclusão

A escolha entre AWS e Google Cloud para implementar Data Science as a Service (DSaaS) integrado com MLOps deve ser orientada não apenas pelo estágio atual da empresa, mas também por suas aspirações futuras e estratégia de crescimento. Startups e pequenas empresas podem se beneficiar do modelo de preços flexível e da variedade de serviços oferecidos pela AWS, facilitando a escalabilidade inicial e a experimentação. Por outro lado, empresas em fase de expansão ou com necessidades avançadas de análise de dados podem encontrar na Google Cloud uma plataforma mais alinhada, especialmente devido às suas capacidades superiores em IA e machine learning.

A decisão final deve levar em conta fatores como a curva de aprendizado das equipes, a compatibilidade com ferramentas e sistemas existentes, e o orçamento disponível. Independentemente da escolha, tanto a AWS quanto a Google Cloud oferecem um ecossistema robusto que pode impulsionar o sucesso de uma iniciativa DSaaS, proporcionando agilidade, eficiência operacional e insights valiosos para a tomada de decisão. A integração com práticas de MLOps dentro dessas plataformas assegura um gerenciamento otimizado do ciclo de vida de modelos de Machine Learning, essencial para manter a competitividade no mercado atual.