

Orientações para elaboração do Trabalho Interdisciplinar Big Data + ML

2º semestre 2022 – Prof. Bianca

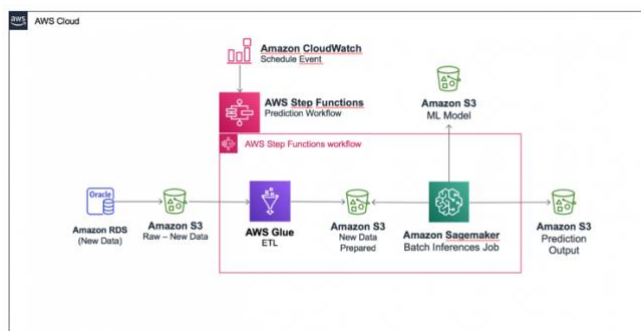
Detalhamento da parte do trabalho referente à Big Data

Processamento de algoritmos de ML em ambiente distribuído, SAGEMAKER.

1. Organização do conteúdo do trabalho

Os trabalhos devem ter os seguintes tópicos:

- Introdução, que deve ser uma breve contextualização.
 - Descrição dos dados (fonte, período, volume, formato, amostras dos dados)
 - Workflow (Diagrama de Arquitetura mostrando as ferramentas e ordem das operações no AWS)
- EX:




- Infraestrutura (Configuração do cluster/instâncias do bloco de anotações como número de nós, tipo de máquinas/instâncias, permissões/role).

Configurações da instância do bloco de anotações

Nome	Status	Tipo da instância do bloco de anotações	Identificador da plataforma
MyNotebook	InService	ml.m4.xlarge	Amazon Linux 2, Jupyter Lab 1 (notebook-ai2-v1)
ARN	Hora de criação	Inferência elástica	Versão mínima do IMDS
arn:aws:sagemaker:us-east-1:251570354991:notebook-instance/mynotebook	Nov 05, 2022 16:12 UTC	-	1
Configuração do ciclo de vida	Última atualização	Tamanho do volume	
arn:aws:sagemaker:us-east-1:251570354991:notebook-instance-lifecycle-config/ml-pipeline-c39669a505452112196471w251570354991	Nov 05, 2022 16:18 UTC	5GB EBS	

- Outras seções do notebook

1. Setup	Configuração da infraestrutura Imports Imagem do pipeline de arquitetura
2. Pipeline de dados	Limpeza e Pré-processamento de dados
3. Treinamento e validação	Seguir as orientações do prof. Samuel
4. Resultados	
5. Clean Up	Remoção dos recursos utilizados que possam gerar custos

Usar como modelo os notebooks dos laboratórios do curso de Machine learning da AWS, incluindo aqueles exemplos disponíveis neste ícone  da barra de ferramentas do jupyter lab, que também estão disponíveis neste [Git](#). Dentre os exemplos de notebook, recomendo especialmente este do [pyspark mnist kmeans](#), que é muito bem organizado.

6. Forma de entrega e apresentação

O trabalho deve ser publicado em um notebook disponibilizado no github. No github organizar pastas com os mesmos nomes do item 1.

Enviar o link para este notebook/GIT via **MOODLE até 21 de novembro**, véspera da data em que o grupo fará uma breve apresentação para a turma.

Quanto à apresentação do trabalho

- Todos os membros do grupo devem estar presentes no momento da apresentação.
- Cada apresentação terá duração de 15 minutos

Critérios de Avaliação da parte específica de Big Data, correspondente a 60% da nota final

Parte Específica de Big Data		
Introdução	0,5	6,0
Dados	0,5	
Workflow de Arquitetura	0,5	
Infraestrutura	1,0	
Setup	1,0	
Pipeline de Dados	1,0	
Clean UP	1,0	
Referências	0,5	
Parte Geral	Parte geral	
Demo	2,0	4,0
Arguição	1,0	
Slides	1,0	

Referências:

1. O Workflow de Arquitetura mostra quais as ferramentas/recursos AWS serão usados e em qual sequência. O site <https://online.visual-paradigm.com/pt/diagrams/templates?search=aws> oferece *templates* para criação de workflows da AWS.
2. Seguem alguns notebooks para inspiração de README (observar mais a forma q o conteúdo, tem as seções e figuras que eu pedi)
 - a) [Async-Inference-Walkthrough.ipynb](#)
 - b) [PySpark K-Means Clustering MNIST.ipynb](#)
3. Alguns outros exemplos bons de notebooks usando EMR (no uso de diagramas de fluxos de dados deixa claro bem claro os passos dos notebooks)
 - a) <https://github.com/Mgosi/Big-Data-Analysis-using-MapReduce-in-Hadoop#readme>
 - b) <https://github.com/faiderfl/Big-Data--architecture-aws-spark#readme>