Tech experts

Semana 7

Sagemaker

- Notebooks
- Studio
- Training Job
- Endpoint
- Batch transform
- Pre processing Jobs
- Model monitoring

Sagemaker Notebooks

- Jupyter Notebook
- Pode acessar de dentro do console AWS, ou de fora, usando endpoint fornecido pela AWS. (isso é possível caso suas configurações de rede permitam)
- Lifecycle configurations: Executa script de configuração sempre que a instacia do seu notebook é inicializada
- Pode ser integrado om github, code commit ou qualquer outro servidor git
- https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/ex1-prepare.html

Sagemaker Training Jobs

- Uma série de algoritmos já previamente disponibilizados, assim como imagens previamende buildadas
- Ex1 (imagem pré-buildada): https://github.com/aws/amazon-sagemaker-examples/blob/default/%20%20%20%20%20%20build_and_train_models/sm-regression_xgboost/sm-regression_xgboost.ipynb
- Ex2 (algoritmo previamente disponibilizado):
- https://github.com/aws/amazon-sagemaker-examples/blob/default/%20%20%20%20%20%20build_and_train_models/sm-random_cut_forest_example/sm-random_cut_forest_example.ipynb

Sagemaker Training Jobs

- Hyperarams tuning Jobs: Busca dos melhores hyperparametros para seu algoritmo
- Ex: https://github.com/aws/amazon-sagemaker-examples/blob/default/%20%20%20%20%20%20build_and_train_models/sm-hyperparameter_tuning_pytorch/sm-hyperparameter_tuning_pytorch.ipynb

Sagemaker Inference

- Endpoint
  - Endpoint de modelo único
  - multi-model endpoint (mesma imagem, múltiplos modelos, único endpoint)
    - https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/create-multi-model-endpoint.html
  - Multi-container endpoit (múltiplas imagens, único endpoint)
    - Invocacao direta:
    - https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/multi-container-create.html
    - Pipeline de inferência:
    - https://github.com/aws/amazon-sagemaker-examples/blob/main/sagemaker-python-sdk/scikit_learn_inference_pipeline/Inference%20Pipeline%20with%20Scikit-learn%20and%20Linear%20Learner.ipynb

Sagemaker Inference

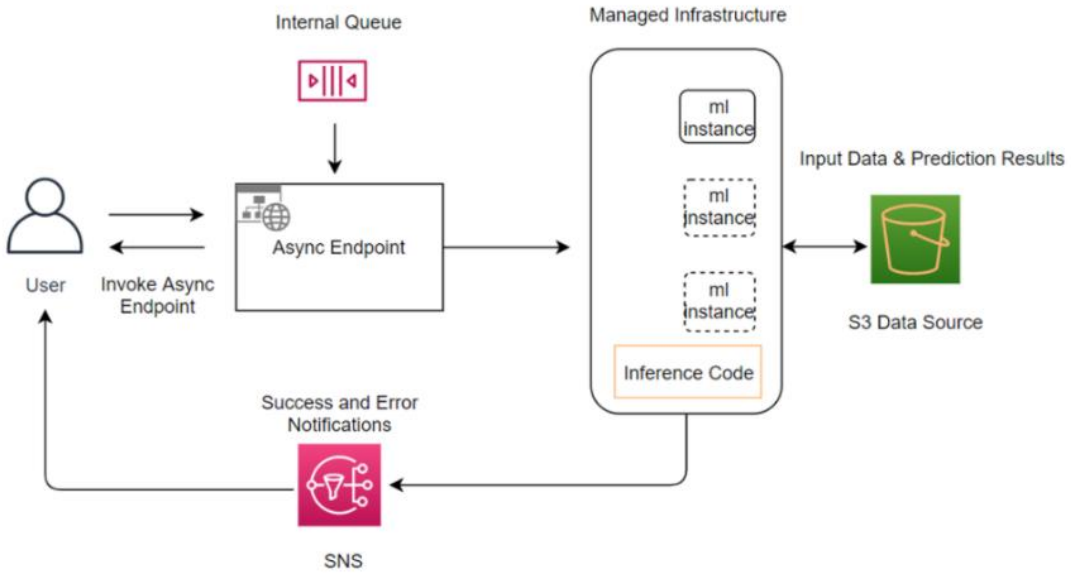
- Endpoint
  - Auto scalling:
    - Dinamica: Acompanha a Métrica do cloudwatch e escala de acordo com Valor Alvo
      - Ex: Número médio de invocações por instancia com valor alvo de 70
    - Baseada em cronograma: realiza a escalabilidade em horários especificos
      - Pode ser configurada em paralelo com a dinamica

Sagemaker Inference

- Endpoint Serverless
  - Conteiners sagemaker ou seus próprios
  - Tamanho máximo da imagem 10GB
  - Memoria de no min 1 GB e no máximo de 6GB (incremental a cada 1GB)
  - Similar a lambda, quanto mais memoria RAM, mais vCPUs
  - Cota de simultaneidade, de ate 200
  - Cold start

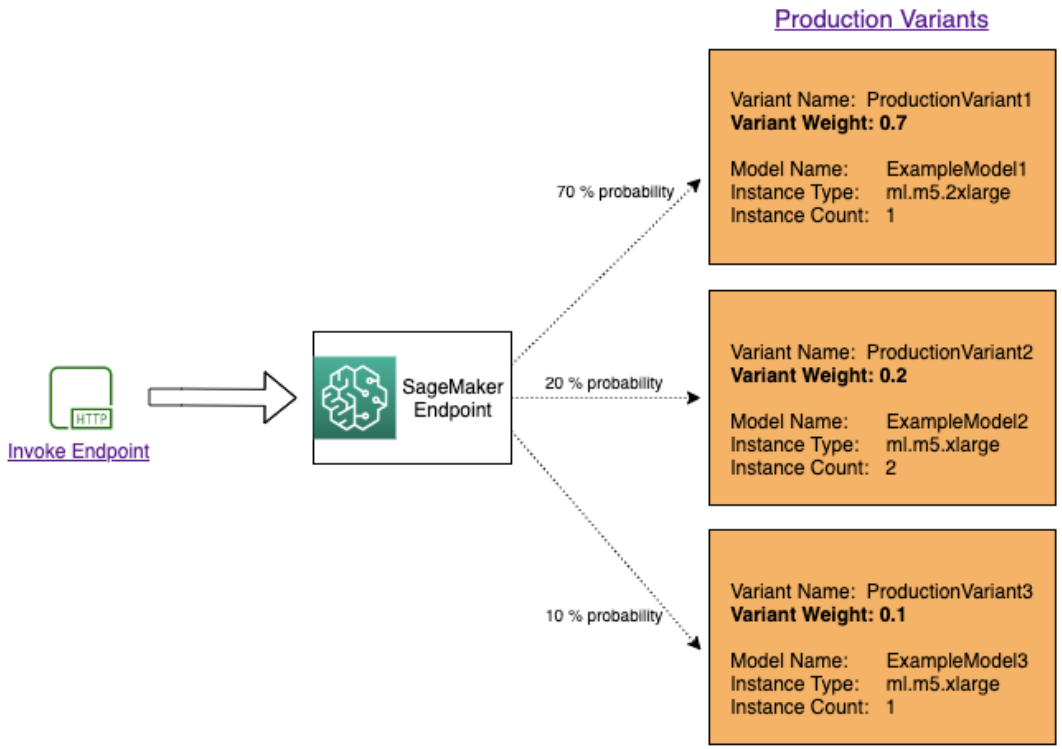
Sagemaker Endpoint Assincrono

- Ideal para cargas grandes (até 1 GB)
- Tempo de processamento de 1h
- Requesitos de latência de quase tempo real
- Escala para 0 (só paga enquanto estiver sendo usado)



Sagemaker Endpoint Testes em produçao

- Distribuicao de trafego
- https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/model-ab-tes



Sagemaker batch transform (transform job)

- Pre processamento
- Inferencia em grandes conjutos de dados
- Inferencia quando não é necessário endpoint
  - Funciona bem com csv e json, não tem suporte nativo para parquet

Sagemaker Monitoring

- Captura de dados
  - Voce pode capturar tanto dados de endpoint quanto de batch transform
  - Captura do batch transform é apenas referência para os dados (manifestos)
- Qualidade de dados (data drift)
  - Baseline
  - Statistics
  - Violacoes
- Model quality
  - Baseline
  - Ground truth
  - Metricas de regressão, classificação binaria, multiclasse
- Sagemaker Clarify
  - Vies (bias): Recomendo fortemente leitura: https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/clarify-model-monitor-bias-drift.html
  - Explainabilty: https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/clarify-model-monitor-feature-attribution-drift.html

Topicos não abordados que podem ser cobrados

  - Sagemaker debugger
  - Sagemaker Experiments
  - Sagemaker Canvas
- Dicas gerais da prova:
  - Para quem for fazer online: ambiente totalmente livre de destraçoes (retire livros, não use duas telas, tire relógio, celular fora do alcance)
  - Voce pode marcar questões para revisar depois, utilize essa função.
  - É comum encontrar respostas para outras questões na própria prova.
  - Voce não precisa saber falar inglês, mas precisa conseguir ler e se comunicar o básico via chat.
  - Antes de agendar a prova, solicite a acomodação para falantes não nativos, são 30 min extras para a prova.
- Boa prova, contem comigo para tirar dúvidas no tems, linkedin, etc!

Q1

- A Machine Learning Specialist is building a model that will perform time series forecasting using Amazon SageMaker. The Specialist has finished training the model and is now planning to perform load testing on the endpoint so they can configure Auto Scaling for the model variant.
  Which approach will allow the Specialist to review the latency, memory utilization, and CPU utilization during the load test?

- A. Review SageMaker logs that have been written to Amazon S3 by leveraging Amazon Athena and Amazon QuickSight to visualize logs as they are being produced.
- B. Generate an Amazon CloudWatch dashboard to create a single view for the latency, memory utilization, and CPU utilization metrics that are outputted by Amazon SageMaker.
- C. Build custom Amazon CloudWatch Logs and then leverage Amazon ES and Kibana to query and visualize the log data as it is generated by Amazon SageMaker.
- D. Send Amazon CloudWatch Logs that were generated by Amazon SageMaker to Amazon ES and use Kibana to query and visualize the log data.

Q1

- A Machine Learning Specialist is building a model that will perform time series forecasting using Amazon SageMaker. The Specialist has finished training the model and is now planning to perform load testing on the endpoint so they can configure Auto Scaling for the model variant.
  Which approach will allow the Specialist to review the latency, memory utilization, and CPU utilization during the load test?

- A. Review SageMaker logs that have been written to Amazon S3 by leveraging Amazon Athena and Amazon QuickSight to visualize logs as they are being produced.
- B. Generate an Amazon CloudWatch dashboard to create a single view for the latency, memory utilization, and CPU utilization metrics that are outputted by Amazon SageMaker.
- C. Build custom Amazon CloudWatch Logs and then leverage Amazon ES and Kibana to query and visualize the log data as it is generated by Amazon SageMaker.
- D. Send Amazon CloudWatch Logs that were generated by Amazon SageMaker to Amazon ES and use Kibana to query and visualize the log data.

Q2

- A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training. The
Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs.
What does the Specialist need to do?

- A. Bundle the NVIDIA drivers with the Docker image.
- B. Build the Docker container to be NVIDIA-Docker compatible.
- C. Organize the Docker container's file structure to execute on GPU instances.
- D. Set the GPU flag in the Amazon SageMaker CreateTrainingJob request body.

Q2

- A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training. The
Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs.
What does the Specialist need to do?

- A. Bundle the NVIDIA drivers with the Docker image.
- B. Build the Docker container to be NVIDIA-Docker compatible.
- C. Organize the Docker container's file structure to execute on GPU instances.
- D. Set the GPU flag in the Amazon SageMaker CreateTrainingJob request body.

Q3

- When submitting Amazon SageMaker training jobs using one of the built-in algorithms, which common parameters MUST be specified? (Choose three.)

- A. The training channel identifying the location of training data on an Amazon S3 bucket.
- B. The validation channel identifying the location of validation data on an Amazon S3 bucket.
- C. The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users.
- D. Hyperparameters in a JSON array as documented for the algorithm used.
- E. The Amazon EC2 instance class specifying whether training will be run using CPU or GPU.
- F. The output path specifying where on an Amazon S3 bucket the trained model will persist

Q3

- When submitting Amazon SageMaker training jobs using one of the built-in algorithms, which common parameters MUST be specified? (Choose three.)

- A. The training channel identifying the location of training data on an Amazon S3 bucket.
- B. The validation channel identifying the location of validation data on an Amazon S3 bucket.
- C. The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users.
- D. Hyperparameters in a JSON array as documented for the algorithm used.
- E. The Amazon EC2 instance class specifying whether training will be run using CPU or GPU.
- F. The output path specifying where on an Amazon S3 bucket the trained model will persist

Q4

- A machine learning (ML) specialist is retraining a new version of a model that is already in production. The model is deployed as an endpoint in Amazon SageMaker. When the retraining is complete, the ML specialist will test the new version of the model before removing the existing production model. The ML specialist's objective is to test the new model at the same time that the existing model is handling the majority of the requests. The deployment must result in minimum disruption to the users of the endpoint.

  Which deployment will meet these requirements with the LEAST operational overhead?
- A: Create a new, separate endpoint in Sagemaker for the newly trained model. Ask a small percentagem of users to call that endpoint instead
- B: Update the endpoint configuration of the production endpoint to include the new model as a ProductionVariant. Set the InitialInstanceCount to 1 and the InstanceType to ml.t2.médium
- C: Create a new, separate endpoint in Sagemaker for the newly trained model. Add another endpoint in Amazon Api Gateway to send the majority of the traffic to the existing endpoint and to send a small percentagem of the traffic to the new endpoint
- D: Update the endpoint configuration of the existing endpoint to include the new model as a Production Variant. Set the InitialVariantWeight for the new model to be a small percentagem of the original ProductionVariant.

Q4

- A machine learning (ML) specialist is retraining a new version of a model that is already in production. The model is deployed as an endpoint in Amazon SageMaker. When the retraining is complete, the ML specialist will test the new version of the model before removing the existing production model. The ML specialist's objective is to test the new model at the same time that the existing model is handling the majority of the requests. The deployment must result in minimum disruption to the users of the endpoint.

  Which deployment will meet these requirements with the LEAST operational overhead?
- A: Create a new, separate endpoint in Sagemaker for the newly trained model. Ask a small percentagem of users to call that endpoint instead
- B: Update the endpoint configuration of the production endpoint to include the new model as a ProductionVariant. Set the InitialInstanceCount to 1 and the InstanceType to ml.t2.médium
- C: Create a new, separate endpoint in Sagemaker for the newly trained model. Add another endpoint in Amazon Api Gateway to send the majority of the traffic to the existing endpoint and to send a small percentagem of the traffic to the new endpoint
- D: Update the endpoint configuration of the existing endpoint to include the new model as a Production Variant. Set the InitialVariantWeight for the new model to be a small percentagem of the original ProductionVariant.

Q5

A machine learning (ML) specialist is building a new recommendation engine. The ML specialist wants to test multiple models by using live data in a beta environment where customers will interact with the model. Based on these interactions, the ML specialist will compare models by using A/B testing. The ML specialist then will select and deploy the best model.
What is the MOST operationally efficient way to test the multiple model variants?

- A: Use Sagemaker to deploy diferente versions of the model behid a single endpoint. Route a percentagem of traffic to each version of the model. Select the best performing model. Reroute 100% of traffic to that model
- B: Deploy multiple versions of the model on EC2 instances behind na Application Load Balance. Evaluate the model performances. Terminate the EC2 instances that host poorly performing models.
- C: Use Sagemaker to deploy na endpoint for each model. Use na Application Load Balance to Route a percentagem off traffic to each model. Gradually Route 100% of traffic to the best model
- D: Use CodePipeline for a blue/green deployment. Use na Apllication Load Balancer to Route a percentagem of traffic to each model. Gradually Route 100% of traffic to the best model.

Q5

A machine learning (ML) specialist is building a new recommendation engine. The ML specialist wants to test multiple models by using live data in a beta environment where customers will interact with the model. Based on these interactions, the ML specialist will compare models by using A/B testing. The ML specialist then will select and deploy the best model.
What is the MOST operationally efficient way to test the multiple model variants?

- A: Use Sagemaker to deploy diferente versions of the model behid a single endpoint. Route a percentagem of traffic to each version of the model. Select the best performing model. Reroute 100% of traffic to that model
- B: Deploy multiple versions of the model on EC2 instances behind na Application Load Balance. Evaluate the model performances. Terminate the EC2 instances that host poorly performing models.
- C: Use Sagemaker to deploy na endpoint for each model. Use na Application Load Balance to Route a percentagem off traffic to each model. Gradually Route 100% of traffic to the best model
- D: Use CodePipeline for a blue/green deployment. Use na Apllication Load Balancer to Route a percentagem of traffic to each model. Gradually Route 100% of traffic to the best model.