Morris & Opazo

aws partner network

Advanced
Consulting Partner

Data & Analytics Competency
Solution Provider
Public Sector Partner
Immersion Day Partner
AWS Lambda

Automated monitoring of data drift in machine learning models

We Love the Cloud

morrisopazo.com

contacto@morrisopazo.com

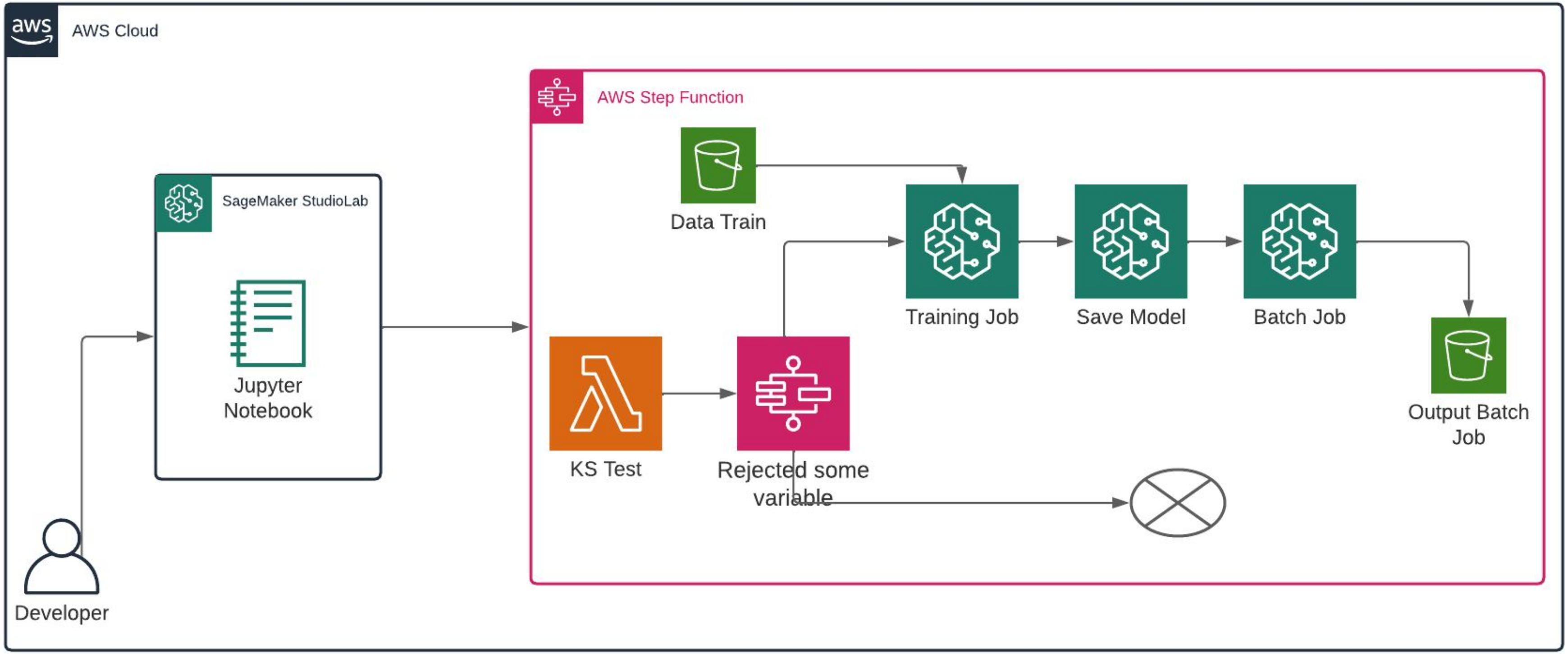
SUMARY

Data drift is a phenomenon in which the input variables of a model change their usual behavior, generating a decrease in the performance of the model. Therefore, generating processes for monitoring potential changes in the population is key when deploying a model in a productive environment.

In the following whitepaper an automated workflow is developed, through step functions, where a solution is implemented that compares the distributions of each variable, through the non-parametric Kolmogorov-Smirnov statistical test, between the original data set and the sample of new data. In case of finding a statistical variation in at least one variable, the model will be updated, training the algorithm, where new predictions are subsequently generated through a batch process.

ARCHITECTURE DIAGRAM

Morris & Opazo



ARCHITECTURE DIAGRAM

AWS services are used for what purposes in this architecture:

- **Sagemaker StudioLab**: it is used to test the code to be executed and perform the tests.

- **Step Functions:** It will be our conductor, indicating the flow step by step as it should follow

- **SageMaker**: will be in charge of processing our work related to Machine Learning, it is the service par excellence when it comes to this. It has an extensive range of instances that will fit the exact requirement.

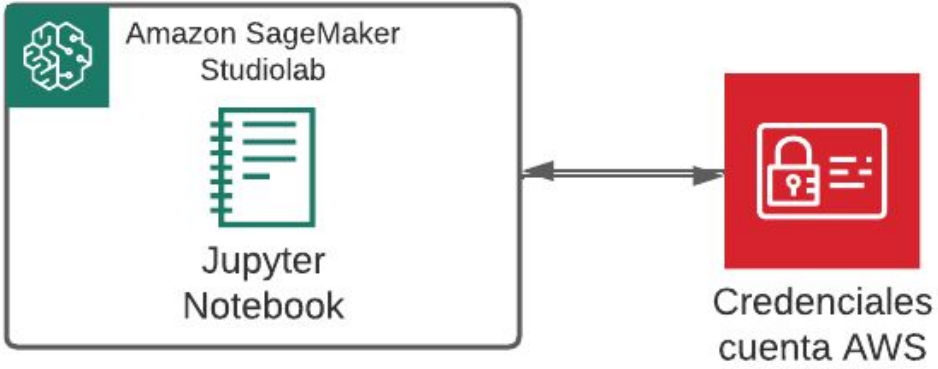
ARCHITECTURE DIAGRAM

- **S3 Bucket:** it is used to obtain the training data and then store the batch results

- **Lambda**: it is used to carry out the test and later its output, it will be our variable to compare

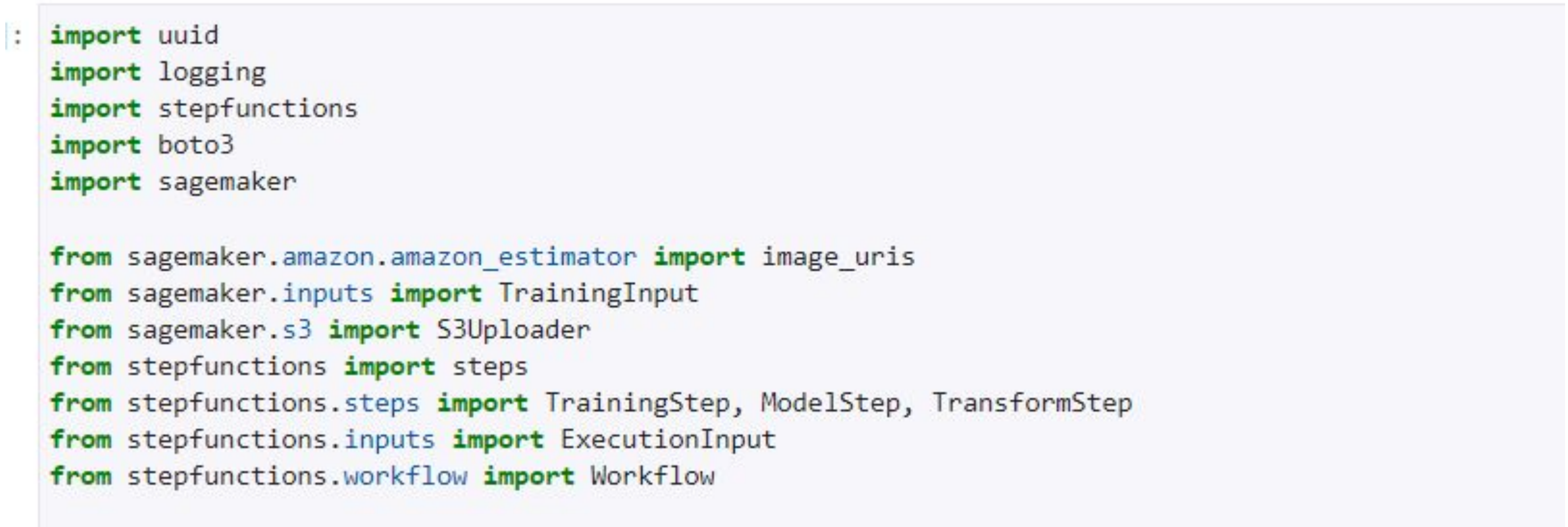
IMPLEMENTATION

1- Associate accounts: Our first step will always be to be aware of the roles that we will need to execute the tasks related to the processes. Therefore, it is vital to have an AWS account to be able to feed the necessary roles to the notebook.



IMPLEMENTATION

2- Load Amazon Sagemaker Studio Labs libraries: Within the modules that will be imported, you will find our main development tool, it will be the step functions data science SDK. It will allow us to create the entire flow within our notebook.

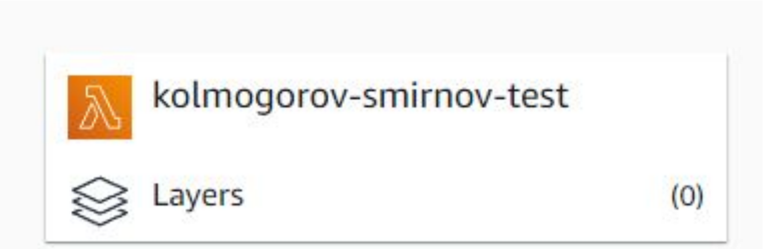


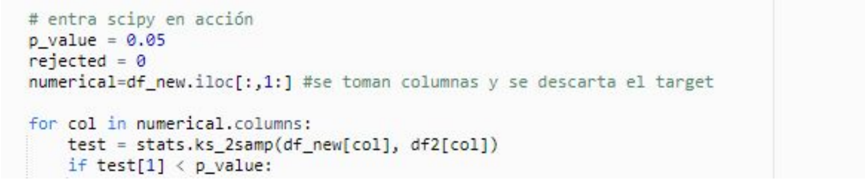
```
import uuid
import logging
import stepfunctions
import boto3
import sagemaker

from sagemaker.amazon.amazon_estimator import image_uris
from sagemaker.inputs import TrainingInput
from sagemaker.s3 import S3Uploader
from stepfunctions import steps
from stepfunctions.steps import TrainingStep, ModelStep, TransformStep
from stepfunctions.inputs import ExecutionInput
from stepfunctions.workflow import Workflow
```

IMPLEMENTATION

3- Lambda function creation: A lambda function is created, which allows applying the Kolmogorov-Smirnov (K-S) statistical test, a non-parametric test, which aims to test the hypothesis that two probability distributions are identical, or failing that , different. The implemented function performs a hypothesis test on each pair of variables associated with the data set with which the model was trained, and the data set with new samples collected. The K-S test is already developed in the **Scipy** python module.





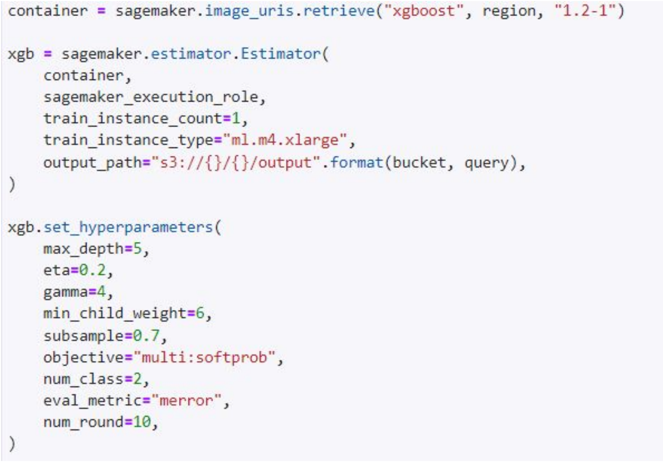
```
# entra scipy en acción
p_value = 0.05
rejected = 0
numerical=df_new.iloc[:,1:] #se toman columnas y se descarta el target

for col in numerical.columns:
    test = stats.ks_2samp(df_new[col], df2[col])
    if test[1] < p_value:
```

IMPLEMENTATION

4- Training configuration: The parameters to execute the training are imposed from the notebook, these help us to develop an optimal model.

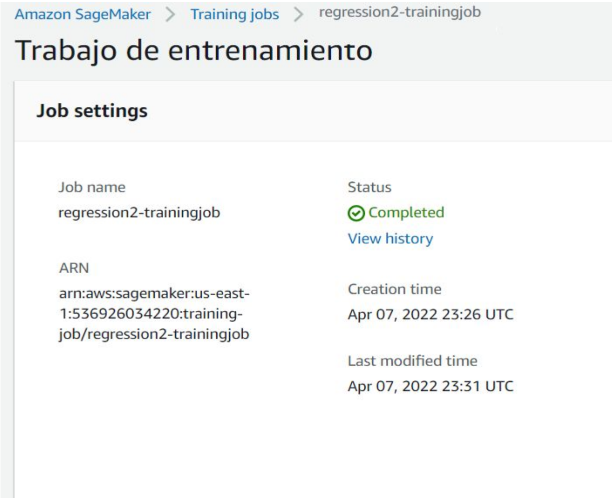
This work will finally be executed from AWS, using a machine optimized for Machine Learning.



```
container = sagemaker.image_uris.retrieve("xgboost", region, "1.2-1")

xgb = sagemaker.estimator.Estimator(
    container,
    sagemaker_execution_role,
    train_instance_count=1,
    train_instance_type="ml.m4.xlarge",
    output_path="s3://{}/{}/output".format(bucket, query),
)

xgb.set_hyperparameters(
    max_depth=5,
    eta=0.2,
    gamma=4,
    min_child_weight=6,
    subsample=0.7,
    objective="multi:softprob",
    num_class=2,
    eval_metric="merror",
    num_round=10,
)
```



Amazon SageMaker > Training jobs > regression2-trainingjob

Trabajo de entrenamiento

Job settings

Job name
regression2-trainingjob

ARN
arn:aws:sagemaker:us-east-1:536926034220:training-job/regression2-trainingjob

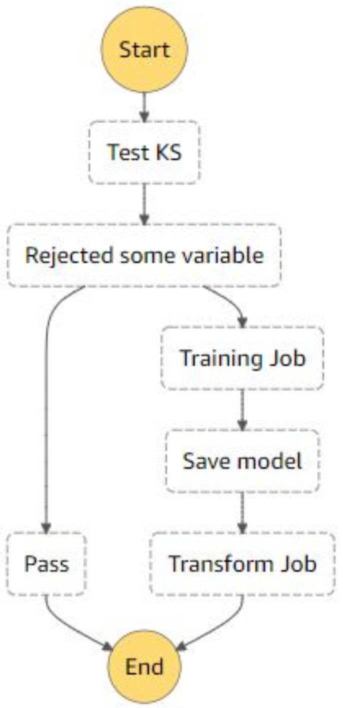
Status
✓ Completed
View history

Creation time
Apr 07, 2022 23:26 UTC

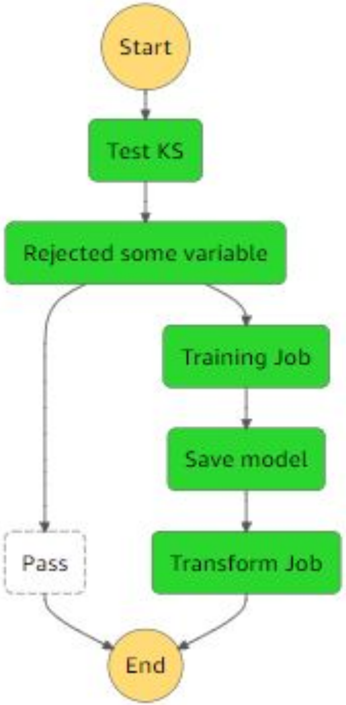
Last modified time
Apr 07, 2022 23:31 UTC

IMPLEMENTATION

5- Flow definition: Once each step of our flow is configured, we generate a pipeline of each stage through a step function. The chain of processes begins by applying the K-S test, explained above, where if there is 1 or more rejected hypothesis tests, the model update will start, launching a training and later an inference job in Batch.



IMPLEMENTATION



6- Step functions flow display: Finally the flow is displayed, where one of the variables has a different distribution in relation to the initial data set, thus generating a rejection in the hypothesis test of the variable, triggering the processes corresponding to the pipeline exposed in the figure.

CONCLUSION

In this exercise, it was possible to implement a tool to monitor possible changes in the population, allowing a new model to be retrained in an automated manner, in order to ensure that the predictions are always correct. For this, the main development was the construction of a Lambda function, adding the necessary configurations to perform statistical tests. Along these same lines, a next step is the incorporation of other types of tools to manage data drift, such as the Population Stability Index (PSI) concept.

RELATED LINKS

Morris & Opazo

- https://docs.aws.amazon.com/sagemaker/latest/dg/studio-lab.html
- https://docs.aws.amazon.com/step-functions/latest/dg/concepts-python-sdk.html
- https://docs.aws.amazon.com/lambda/latest/dg/lambda-python.html
- https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html
- https://docs.aws.amazon.com/sagemaker/latest/dg/batch-transform.html

Morris & Opazo

aws partner network

Advanced

Consulting Partner

- Data & Analytics Competency
- Solution Provider
- Public Sector Partner
- Immersion Day Partner
- AWS Lambda

Expertos en Tecnología de la Nube

We Love the Cloud

Morrisopazo.com

contacto@Morrisopazo.com