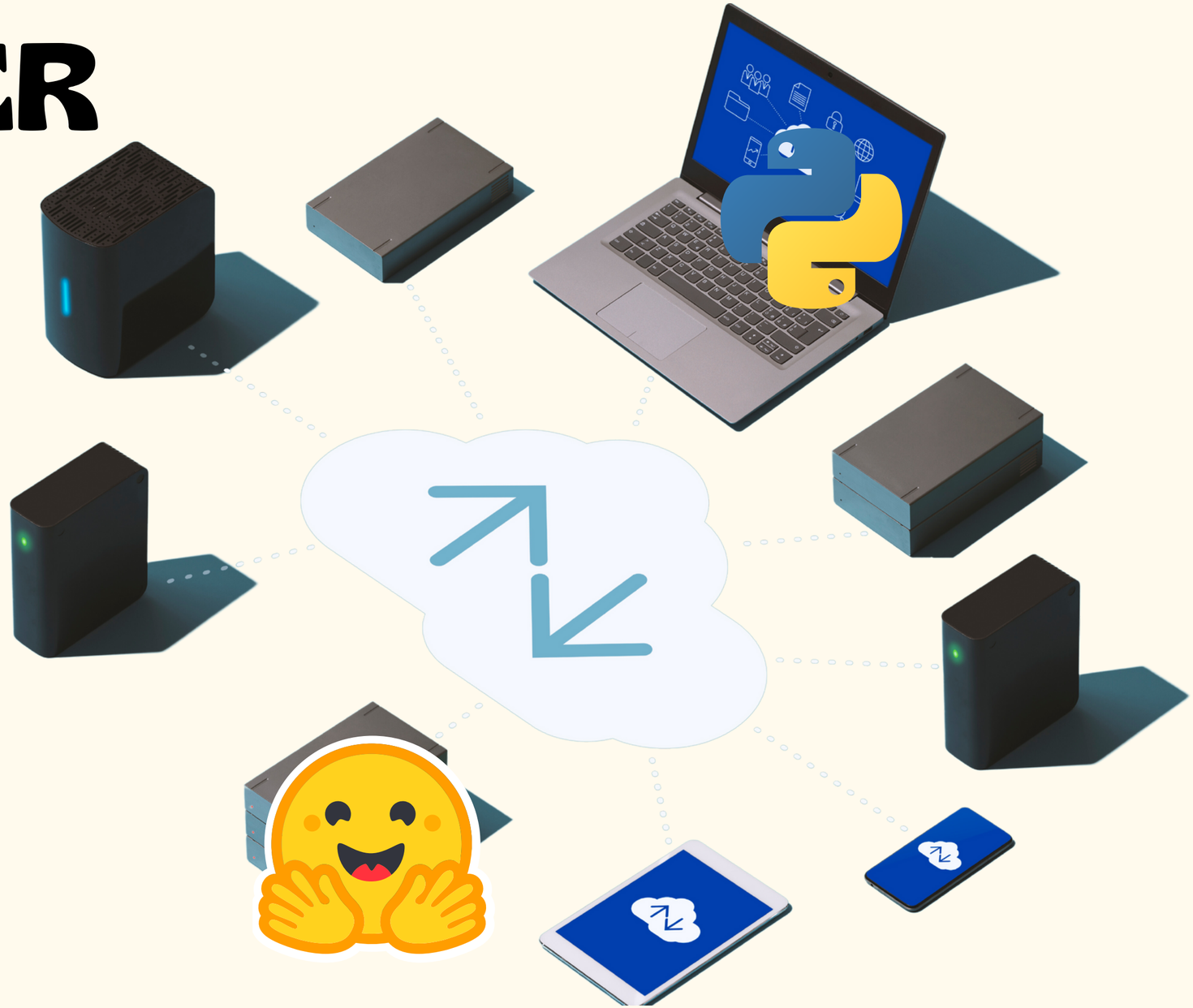
DEPLOY HUGGINGFACE MODELS IN SAGEMAKER

INTRODUCING AWS SAGEMAKER



HTTPS://GITHUB.COM/INSIGHTBUILDER

WHAT CHALLENGE SAGEMAKER SOLVES & HOW

LABEL DATA	**SAGEMAKER STUDIO:** LETS YOU BUILD, TRAIN, DEBUG, DEPLOY, AND MONITOR YOUR MACHINE LEARNING MODELS.
BUILD	SAGEMAKER **NOTEBOOK** INSTANCES: PREPARE, PROCESS DATA, TRAIN & DEPLOY MACHINE LEARNING MODELS FROM A COMPUTE INSTANCE RUNNING THE JUPYTER NOTEBOOK APPLICATION. (VERY SIMILAR TO COLAB ENVIRONMENT)
TRAIN	SAGEMAKER **STUDIO LAB**: STUDIO LAB **IS A FREE SERVICE THAT GIVES YOU ACCESS TO AWS COMPUTE RESOURCES,** IN AN ENVIRONMENT BASED ON OPEN-SOURCE JUPYTERLAB, WITHOUT REQUIRING AN AWS ACCOUNT.
TUNE	**SAGEMAKER CANVAS**: GIVES YOU THE ABILITY TO USE MACHINE LEARNING TO GENERATE PREDICTIONS WITHOUT NEEDING TO CODE.
DEPLOY	SAGEMAKER **GEOSPATIAL**: GIVES YOU THE ABILITY TO BUILD, TRAIN, AND DEPLOY GEOSPATIAL MODELS.
DISCOVER	**RSTUDIO**: RSTUDIO IS AN IDE FOR R, WITH A CONSOLE, SYNTAX-HIGHLIGHTING EDITOR THAT SUPPORTS DIRECT CODE EXECUTION, AND TOOLS FOR PLOTTING, HISTORY, DEBUGGING AND WORKSPACE MANAGEMENT.

STEPS TO DEPLOY THE MODELS

8 STEPS:

1. CREATE ROLE
2. CREATE DOMAIN
3. CREATE USER
4. CREATE STUDIO INSTANCE
5. UNDERSTAND SAGEMAKER CLASSES
6. PULL THE MODEL & STORE IN S3
7. CREATE INFERENCE END POINT
8. PREDICT

CONNECTED WITH:

S3 BUCKETS,
HUGGING FACE HUB,
GIT REPOSITORIES
LINUX USERS

ENVIRONMENT:

DOMAIN,
USERPROFILE
SHARED SPACE
APP

MODELS ACTIVITIES:

SAGEMAKER STUDIO,
SAGEMAKER STUDIO NOTEBOOKS,
RSTUDIO

HTTPS://GITHUB.COM/INSIGHTBUILDER

OVERVIEW ON SAGEMAKER SDK

APIS

- FEATURE STORE APIS
- TRAINING APIS
- DISTRIBUTED TRAINING APIS
- INFERENCE APIS
- GOVERNANCE APIS
- UTILITY APIS

BUILT-IN ALGORITHMS

- AMAZON ESTIMATORS
- TABULAR
- TEXT
- TIME-SERIES
- UNSUPERVISED
- VISION

FRAMEWORKS

1. APACHE MXNET
2. CHAINER
3. HUGGING FACE
4. PYTORCH
5. REINFORCEMENT LEARNING
6. SCIKIT-LEARN
7. SPARKML SERVING
8. TENSORFLOW
9. XGBOOST
10. DEEP JAVA LIBRARY (DJL)

HTTPS://GITHUB.COM/INSIGHTBUILDER

STORING MODEL IN S3

2 MODEL STORAGE OPTION:

STORE IN S3 BUCKET

PULL FROM HUGGINGFACE

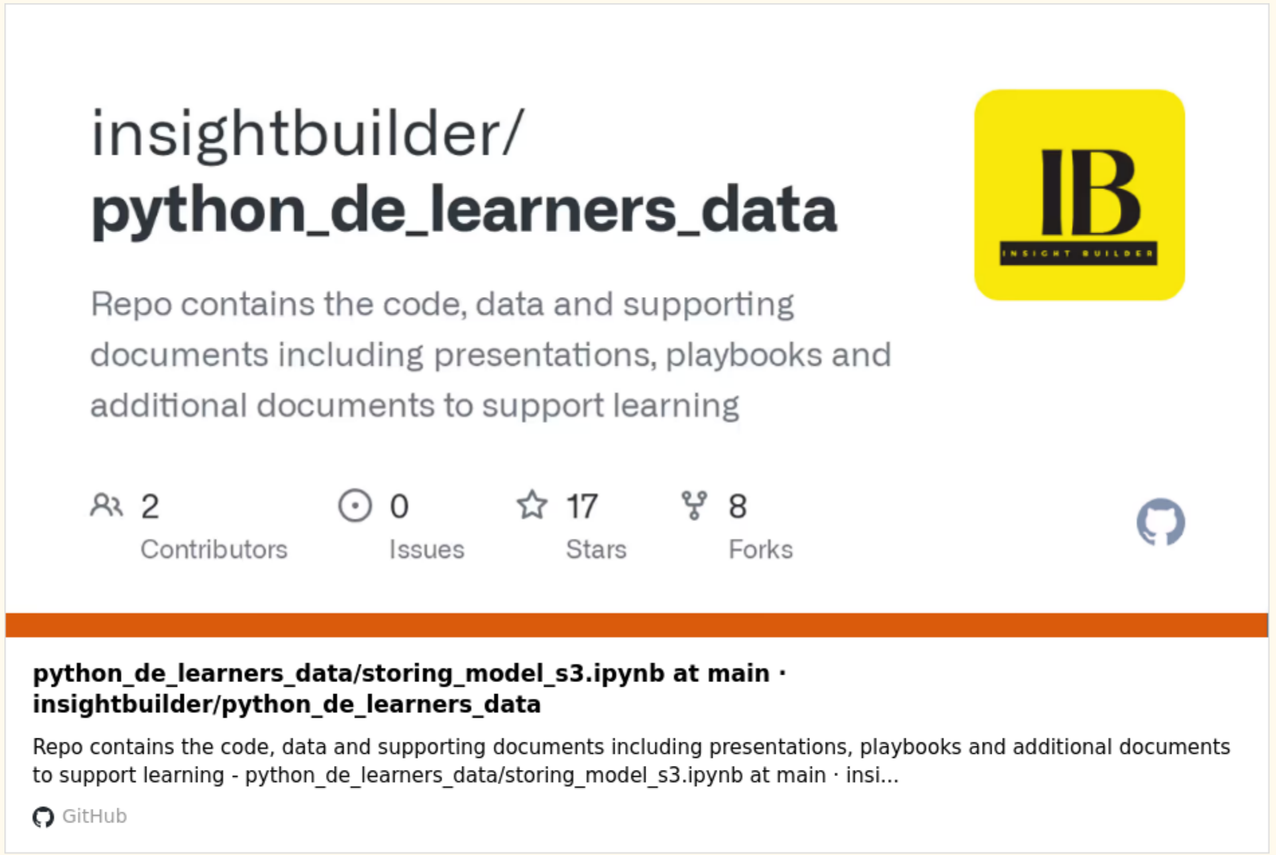
```
PUBLIC S3 URI TO GPT-J ARTIFACT
MODEL_URI="S3://HUGGINGFACE-SAGEMAKER-
MODELS/TRANSFORMERS/4.12.3/PYTORCH/1.9.1/GPT-J/MODEL.TAR.GZ"


HUB = {
 'HF_MODEL_ID':'ELEUTHERAI/GPT-J-6B',
 'HF_TASK':'TEXT-GENERATION'
}
```

IMPORTANT LINKS

- HTTPS://SAGEMAKER.READTHEDOCS.IO/EN/STABLE/FRAMEWORKS/HUGGINGFACE/SAGEMAKER.HUGGINGFACE.HTML
- HTTPS://GITHUB.COM/HUGGINGFACE/NOTEBOOKS/BLOB/MAIN/SAGEMAKER/11_DEPLOY_MODEL_FROM_HF_HUB/DEPLOY_TRANSFORMER_MODEL_FROM_HF_HUB.IPYNB
- HTTPS://GITHUB.COM/HUGGINGFACE/NOTEBOOKS/BLOB/MAIN/SAGEMAKER/10_DEPLOY_MODEL_FROM_S3/DEPLOY_TRANSFORMER_MODEL_FROM_S3.IPYNB
- HTTPS://HUGGINGFACE.CO/ELEUTHERAI/GPT-J-6B
- HTTPS://HUGGINGFACE.CO/DOCS/SAGEMAKER/INDEX

LETS HEAD TO GIT REPO

NOW WE ARE TALKING. PRACTICE???



THANKS FOR WATCHING

HTTPS://GITHUB.COM/INSIGHTBUILDER