

## Cloud Computing Assignment on SageMaker

Name: Vinnamala Sai Sujith

Roll No: 2105339

Section: cse-28

```
[ ]: !pip install transformers "datasets[s3]==2.18.0"
"sagemaker>=2.190.0" _
    ↪"huggingface_hub[cli]" --upgrade
```

```
Requirement already satisfied: transformers in
/usr/local/lib/python3.10/distpackages (4.38.2) Collecting
transformers
  Downloading transformers-4.39.1-py3-none-any.whl (8.8 MB)
    8.8/8.8 MB
  18.6 MB/s eta 0:00:00
Collecting datasets[s3]==2.18.0
  Downloading datasets-2.18.0-py3-none-any.whl (510 kB)
    510.5/510.5
  kB 31.3 MB/s eta 0:00:00
Collecting sagemaker>=2.190.0
  Downloading sagemaker-2.214.1-py3-none-any.whl (1.4 MB)
    1.4/1.4 MB
  36.1 MB/s eta 0:00:00
Requirement already satisfied: huggingface_hub[cli] in
/usr/local/lib/python3.10/dist-packages (0.20.3)
Collecting huggingface_hub[cli]
  Downloading huggingface_hub-0.22.1-py3-none-any.whl (388 kB)
    388.6/388.6
  kB 28.4 MB/s eta 0:00:00
Requirement already satisfied: filelock in
/usr/local/lib/python3.10/distpackages (from datasets[s3]==2.18.0)
(3.13.3)
Requirement already satisfied: numpy>=1.17 in
/usr/local/lib/python3.10/distpackages (from datasets[s3]==2.18.0)
(1.25.2) Requirement already satisfied: pyarrow>=12.0.0 in
```

```

/usr/local/lib/python3.10/dist-packages (from datasets[s3]==2.18.0)
(14.0.2) Requirement already satisfied: pyarrow-hotfix in
/usr/local/lib/python3.10/distpackages (from datasets[s3]==2.18.0)
(0.6)
Collecting dill<0.3.9,>=0.3.0 (from datasets[s3]==2.18.0)
  Downloading dill-0.3.8-py3-none-any.whl (116 kB)
    116.3/116.3

kB 12.3 MB/s eta 0:00:00
Requirement already satisfied: pandas in
/usr/local/lib/python3.10/dist-
packages (from datasets[s3]==2.18.0) (1.5.3)
Requirement already satisfied: requests>=2.19.0
in
/usr/local/lib/python3.10/dist-packages (from datasets[s3]==2.18.0)
(2.31.0) Requirement already satisfied: tqdm>=4.62.1 in
/usr/local/lib/python3.10/dist-
packages (from datasets[s3]==2.18.0) (4.66.2)
Collecting xxhash (from datasets[s3]==2.18.0)
  Downloading xxhash-3.4.1-cp310-cp310-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
    194.1/194.1

kB 11.4 MB/s eta 0:00:00
Collecting multiprocessing (from datasets[s3]==2.18.0)
  Downloading multiprocessing-0.70.16-py310-none-any.whl (134 kB)
    134.8/134.8

kB 11.6 MB/s eta 0:00:00
Requirement already satisfied: fsspec[http]<=2024.2.0,>=2023.1.0 in
/usr/local/lib/python3.10/dist-packages (from datasets[s3]==2.18.0)
(2023.6.0) Requirement already satisfied: aiohttp in
/usr/local/lib/python3.10/distpackages (from datasets[s3]==2.18.0)
(3.9.3)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/distpackages (from datasets[s3]==2.18.0)
(24.0)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.10/dist-
packages (from datasets[s3]==2.18.0) (6.0.1)
Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-2024.3.1-py3-none-any.whl
  (29 kB)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers)
(2023.12.25)
Requirement already satisfied: tokenizers<0.19,>=0.14 in

```

```

/usr/local/lib/python3.10/dist-packages (from transformers) (0.15.2)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.4.2)
Requirement already satisfied: attrs<24,>=23.1.0 in
/usr/local/lib/python3.10/dist-packages (from sagemaker>=2.190.0)
(23.2.0)
Collecting boto3<2.0,>=1.33.3 (from sagemaker>=2.190.0)
  Downloading boto3-1.34.72-py3-none-any.whl (139 kB)
    139.3/139.3

kB 3.0 MB/s eta 0:00:00
Requirement already satisfied: cloudpickle==2.2.1 in
/usr/local/lib/python3.10/dist-packages (from sagemaker>=2.190.0)
(2.2.1) Requirement already satisfied: google-pasta in
/usr/local/lib/python3.10/distpackages (from sagemaker>=2.190.0)
(0.2.0)
Requirement already satisfied: protobuf<5.0,>=3.12 in
/usr/local/lib/python3.10/dist-packages (from sagemaker>=2.190.0)
(3.20.3)
Collecting smdebug-rulesconfig==1.0.1 (from sagemaker>=2.190.0)
  Downloading smdebug_rulesconfig-1.0.1-py2.py3-none-any.whl (20 kB)
Collecting importlib-metadata<7.0,>=1.4.0 (from sagemaker>=2.190.0)
  Downloading importlib_metadata-6.11.0-py3-none-any.whl (23 kB)
Collecting pathos (from sagemaker>=2.190.0)
  Downloading pathos-0.3.2-py3-none-any.whl (82 kB)
    82.1/82.1 kB

6.2 MB/s eta 0:00:00
Collecting schema (from sagemaker>=2.190.0)
  Downloading schema-0.7.5-py2.py3-none-any.whl (17 kB)
Requirement already satisfied: jsonschema in
/usr/local/lib/python3.10/distpackages (from sagemaker>=2.190.0)
(4.19.2)
Requirement already satisfied: platformdirs in
/usr/local/lib/python3.10/distpackages (from sagemaker>=2.190.0)
(4.2.0) Requirement already satisfied: tblib<4,>=1.7.0 in
/usr/local/lib/python3.10/dist-packages (from sagemaker>=2.190.0)
(3.0.0)
Requirement already satisfied: urllib3<3.0.0,>=1.26.8 in
/usr/local/lib/python3.10/dist-packages (from sagemaker>=2.190.0)
(2.0.7)
Collecting docker (from sagemaker>=2.190.0)
  Downloading docker-7.0.0-py3-none-any.whl (147 kB)
    147.6/147.6

kB 14.3 MB/s eta 0:00:00
Requirement already satisfied: psutil in
/usr/local/lib/python3.10/dist-

```

```

packages (from sagemaker>=2.190.0) (5.9.5)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface_hub[cli])
(4.10.0)
Collecting InquirerPy==0.3.4 (from huggingface_hub[cli])
  Downloading InquirerPy-0.3.4-py3-none-any.whl (67 kB)
    67.7/67.7 kB
5.3 MB/s eta 0:00:00
Collecting pfzy<0.4.0,>=0.3.1 (from
InquirerPy==0.3.4->huggingface_hub[cli])
  Downloading pfzy-0.3.4-py3-none-any.whl (8.5 kB)
Requirement already satisfied: prompt-toolkit<4.0.0,>=3.0.1 in
/usr/local/lib/python3.10/dist-packages (from
InquirerPy==0.3.4->huggingface_hub[cli]) (3.0.43)
Collecting botocore<1.35.0,>=1.34.72 (from
boto3<2.0,>=1.33.3->sagemaker>=2.190.0)
  Downloading botocore-1.34.72-py3-none-any.whl (12.0 MB)
    12.0/12.0 MB
40.7 MB/s eta 0:00:00
Collecting jmespath<2.0.0,>=0.7.1 (from boto3<2.0,>=1.33.3-
>sagemaker>=2.190.0)
  Downloading jmespath-1.0.1-py3-none-any.whl (20 kB)
Collecting s3transfer<0.11.0,>=0.10.0 (from
boto3<2.0,>=1.33.3->sagemaker>=2.190.0)
  Downloading s3transfer-0.10.1-py3-none-any.whl (82 kB)
    82.2/82.2 kB
6.2 MB/s eta 0:00:00
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.10/dist-packages (from aiohttp-
>datasets[s3]==2.18.0) (1.3.1)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from aiohttp-
>datasets[s3]==2.18.0) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.10/dist-packages (from aiohttp-
>datasets[s3]==2.18.0) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in
/usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets[s3]==2.18.0) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0
in
/usr/local/lib/python3.10/dist-packages (from aiohttp-
>datasets[s3]==2.18.0) (4.0.3)
Requirement already satisfied: zipp>=0.5 in
/usr/local/lib/python3.10/distpackages (from importlib-
metadata<7.0,>=1.4.0->sagemaker>=2.190.0) (3.18.1)

```

```

Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0-
>datasets[s3]==2.18.0) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-
packages (from requests>=2.19.0->datasets[s3]==2.18.0) (3.6)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests>=2.19.0-
>datasets[s3]==2.18.0) (2024.2.2)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-
packages
(from google-pasta->sagemaker>=2.190.0) (1.16.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6
in
/usr/local/lib/python3.10/dist-packages (from jsonschema-
>sagemaker>=2.190.0) (2023.12.1)
Requirement already satisfied: referencing>=0.28.4 in
/usr/local/lib/python3.10/dist-packages (from jsonschema-
>sagemaker>=2.190.0) (0.34.0)
Requirement already satisfied: rpds-py>=0.7.1 in
/usr/local/lib/python3.10/dist-
packages (from jsonschema->sagemaker>=2.190.0) (0.18.0)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from pandas-
>datasets[s3]==2.18.0) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.10/distpackages (from pandas-
>datasets[s3]==2.18.0) (2023.4) Collecting ppft>=1.7.6.8 (from
pathos->sagemaker>=2.190.0)
  Downloading ppft-1.7.6.8-py3-none-any.whl (56 kB)
    56.8/56.8 kB
4.3 MB/s eta 0:00:00
Collecting pox>=0.3.4 (from pathos->sagemaker>=2.190.0)
  Downloading pox-0.3.4-py3-none-any.whl (29 kB)
Collecting aiobotocore<3.0.0,>=2.5.4 (from s3fs-
>datasets[s3]==2.18.0)
  Downloading aiobotocore-2.12.1-py3-none-any.whl (76 kB)
    76.3/76.3 kB
5.0 MB/s eta 0:00:00
INFO: pip is looking at multiple versions of s3fs to determine which
version is compatible with other requirements. This could take a
while.
Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-2024.3.0-py3-none-any.whl (29 kB)
  Downloading s3fs-2024.2.0-py3-none-any.whl (28 kB)
  Downloading s3fs-2023.12.2-py3-none-any.whl (28 kB)

```

```

    Downloading s3fs-2023.12.1-py3-none-any.whl (28 kB)
    Downloading s3fs-2023.10.0-py3-none-any.whl (28 kB)
Collecting aiobotocore~=2.7.0 (from s3fs->datasets[s3]==2.18.0)
    Downloading aiobotocore-2.7.0-py3-none-any.whl (73 kB)
      73.5/73.5 kB
4.0 MB/s eta 0:00:00
Collecting s3fs (from datasets[s3]==2.18.0)
    Downloading s3fs-2023.9.2-py3-none-any.whl (28 kB)
Collecting aiobotocore~=2.5.4 (from s3fs->datasets[s3]==2.18.0)
    Downloading aiobotocore-2.5.4-py3-none-any.whl (73 kB)
      73.4/73.4 kB
3.3 MB/s eta 0:00:00
Collecting s3fs (from datasets[s3]==2.18.0)
    Downloading s3fs-2023.9.1-py3-none-any.whl (28 kB)
INFO: pip is looking at multiple versions of s3fs to determine which
version is compatible with other requirements. This could take a
while.
    Downloading s3fs-2023.9.0-py3-none-any.whl (28 kB)
    Downloading s3fs-2023.6.0-py3-none-any.whl (28 kB)
Requirement already satisfied: contextlib2>=0.5.5 in
/usr/local/lib/python3.10/dist-packages (from schema-
>sagemaker>=2.190.0) (21.6.0)
INFO: pip is looking at multiple versions of aiobotocore to determine
which version is compatible with other requirements. This could take
a while.
Collecting aiobotocore~=2.5.0 (from s3fs->datasets[s3]==2.18.0)
    Downloading aiobotocore-2.5.3-py3-none-any.whl (73 kB)
      73.3/73.3 kB
4.7 MB/s eta 0:00:00
    Downloading aiobotocore-2.5.2-py3-none-any.whl (72 kB)
      72.9/72.9 kB
6.7 MB/s eta 0:00:00
    Downloading aiobotocore-2.5.1-py3-none-any.whl (72 kB)
      72.8/72.8 kB
5.2 MB/s eta 0:00:00
    Downloading aiobotocore-2.5.0-py3-none-any.whl (72 kB)
      72.7/72.7 kB
8.2 MB/s eta 0:00:00
INFO: This is taking longer than usual. You might need to provide
the dependency resolver with stricter constraints to reduce runtime.
See https://pip.pypa.io/warnings/backtracking for guidance. If you
want to abort this run, press Ctrl + C.
Collecting s3fs (from datasets[s3]==2.18.0)
    Downloading s3fs-2023.5.0-py3-none-any.whl (28 kB)
    Downloading s3fs-2023.4.0-py3-none-any.whl (28 kB)
    Downloading s3fs-2023.3.0-py3-none-any.whl (27 kB)

```

```

Collecting aiobotocore~=2.4.2 (from s3fs->datasets[s3]==2.18.0)
  Downloading aiobotocore-2.4.2-py3-none-any.whl (66 kB)
    66.8/66.8 kB
5.7 MB/s eta 0:00:00
Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-2023.1.0-py3-none-any.whl (27 kB)
  Downloading s3fs-2022.11.0-py3-none-any.whl (27 kB)
  Downloading s3fs-2022.10.0-py3-none-any.whl (27 kB)
  Downloading s3fs-2022.8.2-py3-none-any.whl (27 kB)
  Downloading s3fs-2022.8.1-py3-none-any.whl (27 kB)
  Downloading s3fs-2022.8.0-py3-none-any.whl (27 kB)
  Downloading s3fs-2022.7.1-py3-none-any.whl (27 kB)
Collecting aiobotocore~=2.3.4 (from s3fs->datasets[s3]==2.18.0)
  Downloading aiobotocore-2.3.4-py3-none-any.whl (64 kB)
    64.7/64.7 kB
8.1 MB/s eta 0:00:00
Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-2022.7.0-py3-none-any.whl (27 kB)
  Downloading s3fs-2022.5.0-py3-none-any.whl (27 kB)
  Downloading s3fs-2022.3.0-py3-none-any.whl (26 kB)
Collecting aiobotocore~=2.2.0 (from s3fs->datasets[s3]==2.18.0)
  Downloading aiobotocore-2.2.0.tar.gz (59 kB)
    59.7/59.7 kB
7.8 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-2022.2.0-py3-none-any.whl (26 kB)
Collecting aiobotocore~=2.1.0 (from s3fs->datasets[s3]==2.18.0)
  Downloading aiobotocore-2.1.2-py3-none-any.whl (55 kB)
    56.0/56.0 kB
7.2 MB/s eta 0:00:00
Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-2022.1.0-py3-none-any.whl (25 kB)
  Downloading s3fs-2021.11.1-py3-none-any.whl (25 kB)
Collecting aiobotocore~=2.0.1 (from s3fs->datasets[s3]==2.18.0)
  Downloading aiobotocore-2.0.1.tar.gz (54 kB)
    54.5/54.5 kB
7.2 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-2021.11.0-py3-none-any.whl (25 kB)
Collecting aiobotocore~=1.4.1 (from s3fs->datasets[s3]==2.18.0)
  Downloading aiobotocore-1.4.2.tar.gz (52 kB)
    52.5/52.5 kB
6.0 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done

```

```

Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-2021.10.1-py3-none-any.whl (26 kB)
  Downloading s3fs-2021.10.0-py3-none-any.whl (26 kB)
  Downloading s3fs-2021.9.0-py3-none-any.whl (26 kB)
  Downloading s3fs-2021.8.1-py3-none-any.whl (26 kB)
  Downloading s3fs-2021.8.0-py3-none-any.whl (26 kB)
  Downloading s3fs-2021.7.0-py3-none-any.whl (25 kB)
  Downloading s3fs-2021.6.1-py3-none-any.whl (25 kB)
  Downloading s3fs-2021.6.0-py3-none-any.whl (24 kB)
  Downloading s3fs-2021.5.0-py3-none-any.whl (24 kB)
  Downloading s3fs-2021.4.0-py3-none-any.whl (23 kB)
  Downloading s3fs-0.6.0-py3-none-any.whl (23 kB)
Collecting aiobotocore>=1.0.1 (from s3fs->datasets[s3]==2.18.0)
  Downloading aiobotocore-2.12.0-py3-none-any.whl (76 kB)
    76.1/76.1 kB
5.8 MB/s eta 0:00:00
  Downloading aiobotocore-2.11.2-py3-none-any.whl (76 kB)
    76.1/76.1 kB
4.8 MB/s eta 0:00:00
INFO: pip is looking at multiple versions of aiobotocore to
determine which version is compatible with other requirements.
This could take a while.
  Downloading aiobotocore-2.11.1-py3-none-any.whl (76 kB)
    76.1/76.1 kB
4.6 MB/s eta 0:00:00
  Downloading aiobotocore-2.11.0-py3-none-any.whl (76 kB)
    76.1/76.1 kB
9.1 MB/s eta 0:00:00
  Downloading aiobotocore-2.10.0-py3-none-any.whl (75 kB)
    75.9/75.9 kB
9.9 MB/s eta 0:00:00
  Downloading aiobotocore-2.9.1-py3-none-any.whl (75 kB)
    75.8/75.8 kB
9.5 MB/s eta 0:00:00
  Downloading aiobotocore-2.9.0-py3-none-any.whl (75 kB)
    76.0/76.0 kB
10.7 MB/s eta 0:00:00
INFO: This is taking longer than usual. You might need to provide
the dependency resolver with stricter constraints to reduce runtime.
See https://pip.pypa.io/warnings/backtracking for guidance. If you
want to abort this run, press Ctrl + C.
  Downloading aiobotocore-2.8.0-py3-none-any.whl (75 kB)
    75.0/75.0 kB
10.6 MB/s eta 0:00:00
  Downloading aiobotocore-2.6.0-py3-none-any.whl (73 kB)
    73.4/73.4 kB

```



9.8 MB/s eta 0:00:00  
Downloading aiobotocore-2.4.1-py3-none-any.whl (66 kB)  
66.8/66.8 kB

8.7 MB/s eta 0:00:00  
Downloading aiobotocore-2.4.0-py3-none-any.whl (65 kB)  
65.8/65.8 kB

8.0 MB/s eta 0:00:00  
Downloading aiobotocore-2.3.3.tar.gz (65 kB)  
65.7/65.7 kB

8.1 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-2.3.2.tar.gz (104 kB)  
104.8/104.8

kB 9.1 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-2.3.1.tar.gz (65 kB)  
65.3/65.3 kB

8.1 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-2.3.0.tar.gz (65 kB)  
65.1/65.1 kB

7.1 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-2.1.1.tar.gz (57 kB)  
57.5/57.5 kB

6.9 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-2.1.0.tar.gz (54 kB)  
54.6/54.6 kB

5.7 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-2.0.0.tar.gz (52 kB)  
53.0/53.0 kB

5.1 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.4.1.tar.gz (52 kB)  
52.3/52.3 kB

5.6 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.4.0.tar.gz (51 kB)  
51.6/51.6 kB

4.9 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.3.3.tar.gz (50 kB)  
50.6/50.6 kB

6.4 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.3.2.tar.gz (49 kB)  
49.1/49.1 kB

5.9 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.3.1.tar.gz (48 kB)  
48.8/48.8 kB

5.7 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.3.0.tar.gz (48 kB)  
48.2/48.2 kB

5.2 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.2.2.tar.gz (48 kB)  
48.1/48.1 kB

6.2 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.2.1.tar.gz (48 kB)  
48.0/48.0 kB

6.0 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.2.0.tar.gz (47 kB)  
47.3/47.3 kB

5.6 MB/s eta 0:00:00  
Preparing metadata (setup.py) ... done  
Downloading aiobotocore-1.1.2-py3-none-any.whl (45 kB)  
45.1/45.1 kB

5.6 MB/s eta 0:00:00  
Downloading aiobotocore-1.1.1-py3-none-any.whl (45 kB)  
45.0/45.0 kB

5.9 MB/s eta 0:00:00  
Downloading aiobotocore-1.1.0-py3-none-any.whl (43 kB)  
43.7/43.7 kB

5.1 MB/s eta 0:00:00  
Downloading aiobotocore-1.0.7-py3-none-any.whl (42 kB)  
42.9/42.9 kB

5.2 MB/s eta 0:00:00  
Downloading aiobotocore-1.0.6-py3-none-any.whl (42 kB)  
42.1/42.1 kB

5.1 MB/s eta 0:00:00  
Downloading aiobotocore-1.0.5-py3-none-any.whl (42 kB)  
42.1/42.1 kB

5.7 MB/s eta 0:00:00  
Downloading aiobotocore-1.0.4-py3-none-any.whl (41 kB)  
41.6/41.6 kB

```

5.2 MB/s eta 0:00:00
  Downloading aiobotocore-1.0.3-py3-none-any.whl (40 kB)
    40.8/40.8 kB
5.1 MB/s eta 0:00:00
  Downloading aiobotocore-1.0.2-py3-none-any.whl (40 kB)
    40.8/40.8 kB
5.1 MB/s eta 0:00:00
  Downloading aiobotocore-1.0.1-py3-none-any.whl (40 kB) 40.7/40.7
  kB
5.0 MB/s eta 0:00:00
Collecting s3fs (from datasets[s3]==2.18.0)
  Downloading s3fs-0.5.2-py3-none-any.whl (22 kB)
  Downloading s3fs-0.5.1-py3-none-any.whl (21 kB)
  Downloading s3fs-0.5.0-py3-none-any.whl (21 kB)
  Downloading s3fs-0.4.2-py3-none-any.whl (19 kB)
Requirement already satisfied: wcwidth in
/usr/local/lib/python3.10/distpackages (from
prompttoolkit<4.0.0,>=3.0.1->InquirerPy==0.3.4->huggingface_hub[cli])
(0.2.13) Installing collected packages: xxhash, smdebug-rulesconfig,
schema, ppft, pox, pfzy, jmespath, importlib-metadata, dill,
multiprocess, InquirerPy, huggingface_hub, docker, botocore,
s3transfer, s3fs, pathos, transformers, datasets, boto3, sagemaker
Attempting uninstall: importlib-metadata
  Found existing installation: importlib_metadata 7.1.0
  Uninstalling importlib_metadata-7.1.0:
    Successfully uninstalled importlib_metadata-7.1.0
Attempting uninstall: huggingface_hub
  Found existing installation: huggingface-hub 0.20.3
  Uninstalling huggingface-hub-0.20.3:
    Successfully uninstalled huggingface-hub-0.20.3
Attempting uninstall: transformers
  Found existing installation: transformers 4.38.2
  Uninstalling transformers-4.38.2:
    Successfully uninstalled transformers-4.38.2
Successfully installed InquirerPy-0.3.4 boto3-1.34.72 botocore-
1.34.72 datasets-2.18.0 dill-0.3.8 docker-7.0.0 huggingface_hub-
0.22.1 importlib-metadata-6.11.0 jmespath-1.0.1 multiprocess-0.70.16
pathos-0.3.2 pfzy-0.3.4 pox-0.3.4 ppft-1.7.6.8 s3fs-0.4.2 s3transfer-
0.10.1 sagemaker-2.214.1 schema-0.7.5 smdebug-rulesconfig-1.0.1
transformers-4.39.1 xxhash-3.4.1

```

```
[ ]: !huggingface-cli login --token hf_jwOUrKKKVGGoOUrnBQZZQgJTPhfyTHqidWd
```

Token has not been saved to git credential helper. Pass  
`add\_to\_git\_credential=True` if you want to set the git credential as  
well.

Token is valid (permission: read).  
Your token has been saved to /root/.cache/huggingface/token  
Login successful

[ ]: !aws configure

Requirement already satisfied: sagemaker in  
/usr/local/lib/python3.10/dist-packages (2.214.1)  
Requirement already satisfied: attrs<24,>=23.1.0 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (23.2.0)  
Requirement already satisfied: boto3<2.0,>=1.33.3 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (1.34.72)  
Requirement already satisfied: cloudpickle==2.2.1 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (2.2.1)  
Requirement already satisfied: google-pasta in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (0.2.0)  
Requirement already satisfied: numpy<2.0,>=1.9.0 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (1.25.2)  
Requirement already satisfied: protobuf<5.0,>=3.12 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (3.20.3)  
Requirement already satisfied: smdebug-rulesconfig==1.0.1 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (1.0.1)  
Requirement already satisfied: importlib-metadata<7.0,>=1.4.0 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (6.11.0)  
Requirement already satisfied: packaging>=20.0 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (24.0)  
Requirement already satisfied: pandas in  
/usr/local/lib/python3.10/dist-packages  
(from sagemaker) (1.5.3)  
Requirement already satisfied: pathos in  
/usr/local/lib/python3.10/dist-packages  
(from sagemaker) (0.3.2)  
Requirement already satisfied: schema in  
/usr/local/lib/python3.10/dist-packages  
(from sagemaker) (0.7.5)  
Requirement already satisfied: PyYAML~=6.0 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (6.0.1)  
Requirement already satisfied: jsonschema in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (4.19.2)  
Requirement already satisfied: platformdirs in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (4.2.0)  
Requirement already satisfied: tblib<4,>=1.7.0 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (3.0.0)  
Requirement already satisfied: urllib3<3.0.0,>=1.26.8 in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (2.0.7)  
Requirement already satisfied: requests in  
/usr/local/lib/python3.10/dist-packages (from sagemaker) (2.31.0)

Requirement already satisfied: docker in  
 /usr/local/lib/python3.10/dist-packages (from sagemaker) (7.0.0)

Requirement already satisfied: tqdm in  
 /usr/local/lib/python3.10/dist-packages (from sagemaker) (4.66.2)

Requirement already satisfied: psutil in  
 /usr/local/lib/python3.10/dist-packages  
 (from sagemaker) (5.9.5)

Requirement already satisfied: botocore<1.35.0,>=1.34.72 in  
 /usr/local/lib/python3.10/dist-packages (from boto3<2.0,>=1.33.3->sagemaker) (1.34.72)

Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in  
 /usr/local/lib/python3.10/dist-packages (from boto3<2.0,>=1.33.3->sagemaker) (1.0.1)

Requirement already satisfied: s3transfer<0.11.0,>=0.10.0 in  
 /usr/local/lib/python3.10/dist-packages (from boto3<2.0,>=1.33.3->sagemaker) (0.10.1)

Requirement already satisfied: zipp>=0.5 in  
 /usr/local/lib/python3.10/dist-packages (from importlib-metadata<7.0,>=1.4.0->sagemaker) (3.18.1)

Requirement already satisfied: charset-normalizer<4,>=2 in  
 /usr/local/lib/python3.10/dist-packages (from requests->sagemaker) (3.3.2)

Requirement already satisfied: idna<4,>=2.5 in  
 /usr/local/lib/python3.10/dist-packages (from requests->sagemaker) (3.6)

Requirement already satisfied: certifi>=2017.4.17 in  
 /usr/local/lib/python3.10/dist-packages (from requests->sagemaker) (2024.2.2)

Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages  
 (from google-pasta->sagemaker) (1.16.0)

Requirement already satisfied: jsonschema-specifications>=2023.03.6 in  
 /usr/local/lib/python3.10/dist-packages (from jsonschema->sagemaker) (2023.12.1)

Requirement already satisfied: referencing>=0.28.4 in  
 /usr/local/lib/python3.10/dist-packages (from jsonschema->sagemaker) (0.34.0)

Requirement already satisfied: rpds-py>=0.7.1 in  
 /usr/local/lib/python3.10/dist-packages (from jsonschema->sagemaker) (0.18.0)

Requirement already satisfied: python-dateutil>=2.8.1 in  
 /usr/local/lib/python3.10/dist-packages (from pandas->sagemaker) (2.8.2)

Requirement already satisfied: pytz>=2020.1 in  
 /usr/local/lib/python3.10/dist-packages (from pandas->sagemaker) (2023.4)

Requirement already satisfied: ppft>=1.7.6.8 in  
 /usr/local/lib/python3.10/dist-packages (from pathos->sagemaker) (1.7.6.8)

Requirement already satisfied: dill>=0.3.8 in  
 /usr/local/lib/python3.10/distpackages (from pathos->sagemaker)  
 (0.3.8)  
 Requirement already satisfied: pox>=0.3.4 in  
 /usr/local/lib/python3.10/distpackages (from pathos->sagemaker)  
 (0.3.4)  
 Requirement already satisfied: multiprocessing>=0.70.16 in  
 /usr/local/lib/python3.10/dist-packages (from pathos->sagemaker)  
 (0.70.16)  
 Requirement already satisfied: contextlib2>=0.5.5 in  
 /usr/local/lib/python3.10/dist-packages (from schema->sagemaker)  
 (21.6.0)

```
[ ]: import sagemaker
import boto3
sess = sagemaker.Session()
# sagemaker session bucket -> used for uploading data,
models and logs # sagemaker will automatically create this
bucket if it not exists sagemaker_session_bucket=None if
sagemaker_session_bucket is None and sess is not None: # set
to default bucket if a bucket name is not given
sagemaker_session_bucket = sess.default_bucket()

try:
    role = sagemaker.get_execution_role()
except ValueError:
    iam = boto3.client('iam')
    role =

iam.get_role(RoleName='SageMakerTest')['Role']['Arn'] sess

=

sagemaker.Session(default_bucket=sagemaker_session_bucket)

print(f"sagemaker role arn: {role}") print(f"sagemaker
bucket: {sess.default_bucket()}") print(f"sagemaker
session region: {sess.boto_region_name}")
```

WARNING:sagemaker:Couldn't call 'get\_role' to get Role ARN from role  
 name collab to get Role path.

sagemaker role arn:  
 arn:aws:iam::381491942612:role/SageMakerTest sagemaker

bucket: sagemaker-ap-southeast-2-381491942612 sagemaker  
session region: ap-southeast-2

```
[ ]: from datasets import load_dataset

# Convert dataset to OAI messages
system_message = """You are an text to SQL query translator. Users will ask you
↳ questions in English and you will generate a SQL query based on the provided
↳ SCHEMA.
SCHEMA:
{schema}"""

def create_conversation(sample):
    return {
        "messages":
            {"role": "system", "content": system_message.
↳ format(schema=sample["context"])},
            {"role": "user", "content": sample["question"]},
            {"role": "assistant", "content": sample["answer"]}
    ]
}

# Load dataset from the hub
dataset = load_dataset("b-mc2/sql-create-context", split="train")
dataset = dataset.shuffle().select(range(12500))

# Convert dataset to OAI messages
dataset = dataset.map(create_conversation, remove_columns=dataset.
↳ features, batched=False)
# split dataset into 10,000 training samples and 2,500 test samples
dataset = dataset.train_test_split(test_size=2500/12500)

print(dataset["train"][345]["messages"])
```

Downloading readme: 0%| | 0.00/4.43k [00:00<?, ?B/s]

Downloading data: 0%| | 0.00/21.8M [00:00<?, ?B/s]

Generating train split: 0 examples [00:00, ? examples/s]

Map: 0%| | 0/12500 [00:00<?, ? examples/s]

```
[{'content': 'You are an text to SQL query translator. Users will ask
you questions in English and you will generate a SQL query based on
the provided SCHEMA.\nSCHEMA:\nCREATE TABLE table_12962773_4 (no
INTEGER, player VARCHAR)', 'role': 'system'}, {'content': 'What No is
the player Zoran Erceg', 'role':
'user'}, {'content': 'SELECT MIN(no) FROM table_12962773_4 WHERE
player = "Zoran Erceg"', 'role': 'assistant'}]
```

```
[ ]: # save train_dataset to s3 using our SageMaker session
training_input_path =
f's3://{sess.default_bucket()}/datasets/text-to-sql'

# save datasets to s3
dataset["train"].to_json(f"{training_input_path}/train_dataset.json"
, _
    orient="records")
dataset["test"].to_json(f"{training_input_path}/test_dataset.json", _
    orient="records")

print(f"Training data uploaded to:")
print(f"{training_input_path}/train_dataset.json")
print(f"https://s3.console.aws.amazon.com/s3/buckets/{sess.default_bu
cket() }/?
    region={sess.boto_region_name}&prefix={training_input_path.split('
    /', _ 43)[-1]}")
```

Creating json from Arrow 0%| | 0/10 [00:00<?, ?ba/s]

format:

Creating json from Arrow 0%| | 0/3 [00:00<?, ?ba/s]

format:

Training data uploaded to:

s3://sagemaker-ap-southeast-2-381491942612/datasets/text-  
tosql/train\_dataset.json

https://s3.console.aws.amazon.com/s3/buckets/sagemaker-apsoutheast-2-  
381491942612/?region=ap-southeast-2&prefix=datasets/text-to-sql/

```
[ ]: # hyperparameters, which are passed into the training job
hyperparameters = {
    ### SCRIPT PARAMETERS ###
    'dataset_path': '/opt/ml/input/data/training/train_dataset.json', #
    path_
    where sagemaker will save training dataset
    'model_id': "codellama/CodeLlama-7b-hf", # or `mistralai/
    Mistral-7B-v0.1`
    'max_seq_len': 3072, # max sequence length for_
    model and packing of the dataset
    'use_qlora': True, # use QLoRA model
    ### TRAINING PARAMETERS ###
    'num_train_epochs': 3, # number of training epochs
    'per_device_train_batch_size': 1, # batch size per device_
    during training
    'gradient_accumulation_steps': 4, # number of steps before_
    performing a backward/update pass
    'gradient_checkpointing': True, # use gradient_
```



```

↳checkpointing to save memory
'optim': "adamw_torch_fused",          # use fused adamw optimizer
'logging_steps': 10,                  # log every 10 steps
'save_strategy': "epoch",            # save checkpoint every
↳epoch
'learning_rate': 2e-4,                # learning rate, based on
↳QLoRA paper
'bf16': False,                        # use bfloat16 precision
'tf32': True,                         # use tf32 precision
'max_grad_norm': 0.3,                 # max gradient norm based
↳on QLoRA paper
'warmup_ratio': 0.03,                 # warmup ratio based on
↳QLoRA paper
'lr_scheduler_type': "constant",      # use constant learning
↳rate scheduler
'report_to': "tensorboard",           # report metrics to
↳tensorboard
'output_dir': '/tmp/tun',             # Temporary output
↳directory for model checkpoints
'merge_adapters': True,               # merge LoRA adapters into
↳model for easier deployment
}

```

```

[ ]: from sagemaker.huggingface import
HuggingFace

```

```

# define Training Job Name
job_name = f'codellama-7b-hf-text-to-sql-exp1'

# create the Estimator
huggingface_estimator = HuggingFace(
    entry_point      = 'run_sft.py', # train script
    source_dir       = '/content/', # 'https://github.com/philschmid/
↳llm-sagemaker-sample/blob/main/scripts/tr1', # directory which
includes
↳all the files needed for training instance_type =
    'ml.t3.medium', # instances type used for the
↳training job
    instance_count   = 1,          # the number of instances used
↳for training
    max_run          = 2*24*60*60, # maximum runtime in seconds
↳(days * hours * minutes * seconds)
    base_job_name    = job_name,   # the name of the training job

```

```

    role = role, # Iam role used in training job
    ↪to access AWS ressources, e.g. S3

    volume_size = 300, # the size of the EBS volume in GB
    transformers_version = '4.36', # the transformers version used
    ↪in the training job
    pytorch_version = '2.1', # the pytorch_version version
    ↪used in the training job
    py_version = 'py310', # the python version used in the
    ↪training job
    hyperparameters = hyperparameters, # the hyperparameters passed to
    ↪the training job
    disable_output_compression = True, # not compress output to save
    ↪training time and cost
    environment = {
        "HUGGINGFACE_HUB_CACHE": "/tmp/.cache", # set env
    ↪variable to cache models in /tmp
        "HF_TOKEN":
    ↪"hf_jwOUrKKKVGGoOUrnBQZZQgJTPPhfyTHqidWd" # huggingface token to access gated
    ↪models, e.g. llama 2
    },
)

```

```

[ ]: # define a data input dictionary with our uploaded s3 uris
data = {'training': training_input_path}

# starting the train job with our uploaded datasets as input
huggingface_estimator.fit(data, wait=True)

```

```

[ ]: from sagemaker.huggingface import get_huggingface_llm_image_uri

# retrieve the llm image uri
llm_image = get_huggingface_llm_image_uri(
    "huggingface",
    version="1.4.0",
    session=sess,
)

# print ecr image uri
print(f"llm image uri: {llm_image}")

```

```

llm image uri: 763104351884.dkr.ecr.ap-southeast-
2.amazonaws.com/huggingfacepytorch-tgi-inference:2.1.1-tgi1.4.0-gpu-
py310-cul21-ubuntu20.04

```

```

[ ]: import json from sagemaker.huggingface
import HuggingFaceModel

```

```
# s3 path where the model will be uploaded  
# if you try to deploy the model to a different time add the s3 path here  
model_s3_path = huggingface_estimator.model_data["S3DataSource"]["S3Uri"]
```

```

# sagemaker config
instance_type = "ml.g5.2xlarge"
number_of_gpu = 1
health_check_timeout = 300

# Define Model and Endpoint configuration parameter
config = {
    'HF_MODEL_ID': "/opt/ml/model", # path to where sagemaker stores the model
    'SM_NUM_GPUS': json.dumps(number_of_gpu), # Number of GPU used per replica
    'MAX_INPUT_LENGTH': json.dumps(1024), # Max length of input text
    'MAX_TOTAL_TOKENS': json.dumps(2048), # Max length of the generation _
    ↪(including input text)
}

# create HuggingFaceModel with the image uri
llm_model = HuggingFaceModel(
    role=role,
    image_uri=llm_image,
    model_data={'S3DataSource':{'S3Uri': model_s3_path, 'S3DataType': _
    ↪'S3Prefix', 'CompressionType': 'None'}}},
    env=config
)

```

```

[ ]: # Deploy model to an endpoint
# https://sagemaker.readthedocs.io/en/stable/api/inference/model.html#sagemaker.
    ↪model.Model.deploy
llm = llm_model.deploy(
    initial_instance_count=1,
    instance_type=instance_type,
    container_startup_health_check_timeout=health_check_timeout, # 10 minutes to _
    ↪give SageMaker the time to download the model
)

```

```

[ ]: from transformers import AutoTokenizer
from sagemaker.s3 import S3Downloader

# Load the tokenizer
tokenizer = AutoTokenizer.from_pretrained("codellama/CodeLlama-7b-hf")

# Load the test dataset from s3
S3Downloader.download(f"{training_input_path}/test_dataset.json", ".")
test_dataset = load_dataset("json", data_files="test_dataset.
    ↪json", split="train")
random_sample = test_dataset[345]

```

```

def request(sample):
    prompt = tokenizer.apply_chat_template(sample, tokenize=False,
    ↪add_generation_prompt=True)
    outputs = llm.predict({
        "inputs": prompt,
        "parameters": {
            "max_new_tokens": 512,
            "do_sample": False,
            "return_full_text": False,
            "stop": ["<|im_end|>"],
        }
    })
    return {"role": "assistant", "content": outputs[0]["generated_text"]}.
    ↪strip()

print(random_sample["messages"][1])
request(random_sample["messages"][:2])

```

```

[ ]: from transformers import AutoTokenizer
from sagemaker.s3 import S3Downloader

# Load the tokenizer
tokenizer = AutoTokenizer.from_pretrained("codellama/CodeLlama-7b-hf")

# Load the test dataset from s3
S3Downloader.download(f"{training_input_path}/test_dataset.json", ".")
test_dataset = load_dataset("json", data_files="test_dataset.
    ↪json", split="train")
random_sample = test_dataset[345]

def request(sample):
    prompt = tokenizer.apply_chat_template(sample, tokenize=False,
    ↪add_generation_prompt=True)
    outputs = llm.predict({
        "inputs": prompt,
        "parameters": {
            "max_new_tokens": 512,
            "do_sample": False,
            "return_full_text": False,
            "stop": ["<|im_end|>"],
        }
    })
    return {"role": "assistant", "content": outputs[0]["generated_text"]}.
    ↪strip()

print(random_sample["messages"][1])
request(random_sample["messages"][:2])

```

```
[ ]: from tqdm import tqdm

def evaluate(sample):
    predicted_answer = request(sample["messages"][:2])
    if predicted_answer["content"] == sample["messages"][2]["content"]:
        return 1
    else:
        return 0

success_rate = []
number_of_eval_samples = 1000
# iterate over eval dataset and predict
for s in tqdm(test_dataset.shuffle().select(range(number_of_eval_samples))):
    success_rate.append(evaluate(s))

# compute accuracy
accuracy = sum(success_rate)/len(success_rate)

print(f"Accuracy: {accuracy*100:.2f}%")
```

```
[ ]: llm.delete_model()
llm.delete_endpoint()
```