from pyspark.sql import SparkSession

from pyspark.sql.functions import col

from pathlib import Path

# Create a SparkSession

spark = SparkSession.builder.appName("MyApp").getOrCreate()

# MOA

data\_dir\_moa = Path('/user/anupriya7996/project/input/')

data\_dir\_moa = Path(‘/user/anupriya7996/project/input/Data/mechanismOfAction/’)

moa\_full\_df = spark.read.parquet(str(data\_dir\_moa) + '/\*.parquet')

# AE

data\_dir = Path(‘/user/anupriya7996/project/input/Data/significantAdverseDrugReactions/’)

ae\_full\_df = spark.read.parquet(str(data\_dir) + '/\*.parquet')

# TARGETS-PATHWAY

data\_dir\_targets = Path(‘/user/anupriya7996/project/input/Data/Targets/’)

targets\_full\_df = spark.read.parquet(str(data\_dir\_targets) + '/\*.parquet')

# Show the schema of the loaded DataFrames

moa\_full\_df.printSchema()

ae\_full\_df.printSchema()

targets\_full\_df.printSchema()

# Perform operations on the DataFrames using PySpark APIs

# For example, you can filter or transform columns using functions from the 'pyspark.sql.functions' module.

# For instance, to filter the 'moa\_full\_df' DataFrame based on a condition:

filtered\_moa\_df = moa\_full\_df.filter(col("columnName") > value)

# Do similar operations for 'ae\_full\_df' and 'targets\_full\_df' if needed.

# Stop the SparkSession when you are done

spark.stop()

‘/user/anupriya7996/project/input/Data/mechanismOfAction/’

data\_dir\_target\_ae = Path(‘/user/anupriya7996/project/input/Data/targets/’)

target\_ae\_full\_df = spark.read.parquet(str(data\_dir\_target\_ae))

# Indications

data\_dir\_indications = Path(‘/user/anupriya7996/project/input/Data/indication/’)

indications\_full\_df = spark.read.parquet(str(data\_dir\_indications))

# Show the schema of the loaded DataFrames

target\_ae\_full\_df.printSchema()

indications\_full\_df.printSchema()