Spark Assignment Peer Review

Dhruv SinglaApproach:

Created two files query.py and store_data.py. store_data.py is for loading data and Query.py file to answers the questions in the assignment.

Store_data.py

- get_data_from_url : this function calls the api and returns the result.
- store_data_into_csv : this function will store data in a csv file. The file will be used to get data instead of calling the api again and again.
- store_into_spark_dataframe : this function creates a SparkSession and create a Spark dataframe by excluding the header row and passing the remaining rows as the data to the spark.createDataFrame function. The toDF method is used to assign column names to the dataframe and finally it is returning dataframe.

Query.py

most_affected_state : The function first creates a new column called "death_to_total" by dividing the "death" column by the "total" column using the withColumn method. Then, it orders the rows of the dataframe by the "death_to_total" column in descending order using the orderBy method. The first method is used to retrieve the first row of the ordered dataframe, which will be the row with the highest "death_to_total" value.
Other query functions are similar.

Amit Shukla Approach:

Created two files data.py and app.py. Data.py is for loading data and writing query functions. App.py is just containing the flask code which calls the functions in data.py, no logic for queries is written in app.py

Data.py:

In data.py created class Data.

- __init__ : it calls create_dataframe

- create_dataframe : it first creates a sparkSession. Then defines a schema for the spark dataframe using StructType and StructField, eg - StructType([StructField('SNo',IntegerType())]). Then calls clean_dataset and finally spark's builtin function to create a spark dataframe.
- load_dataset : this is called by clean_dataset. It just returns the dictionary got from the api call.
- remove_stars : called by clean_dataset. It removes * character from state column's values.
- clean_dataset : after getting data from load_dataset it checks in the values of dictionary. The values should be dictionary with state column non-empty, otherwise they are discarded. For each value performs remove_stars on it and then appends the values of the dictionary in a list.\

Query Functions:

- affected - Creates a new column in the dataframe named Affected by Death/Total_Cases(dividing two columns)
- most_affected - Orders the dataframe based on Affected and then selects the first row of State column Rest are similar.

App.py:

It imports the Data class from data.py

- covid_cases : it's the home page and contains hyperlinks to all other pages
- get_most_affected : calls the most_affected from Data Other query functions are similar