Abbiamo cominciato scaricando pymongo per interagire con mongo tramite python.

Installato pyspark per interagire con spark.

Installato spark (con bestemmie su windows),

WINDOWS ONLY: bin hadoop (versione di spark) winutils.dll aggiunto al path (a mano), cartella spark aggiunta al path (a mano)

Creata sessione spark nel codice, aggiungendo il link al db in input e quello in output e specificando che deve scaricare e importare il modulo del connettore da spark a mongo.

Caricati i dati dal database come Dataframe spark. Dataframe è una struttura, wrapper di RDD che ha le proprietà da controllare:

1. Partizionabile
2. Località (se in memoria e tutta)