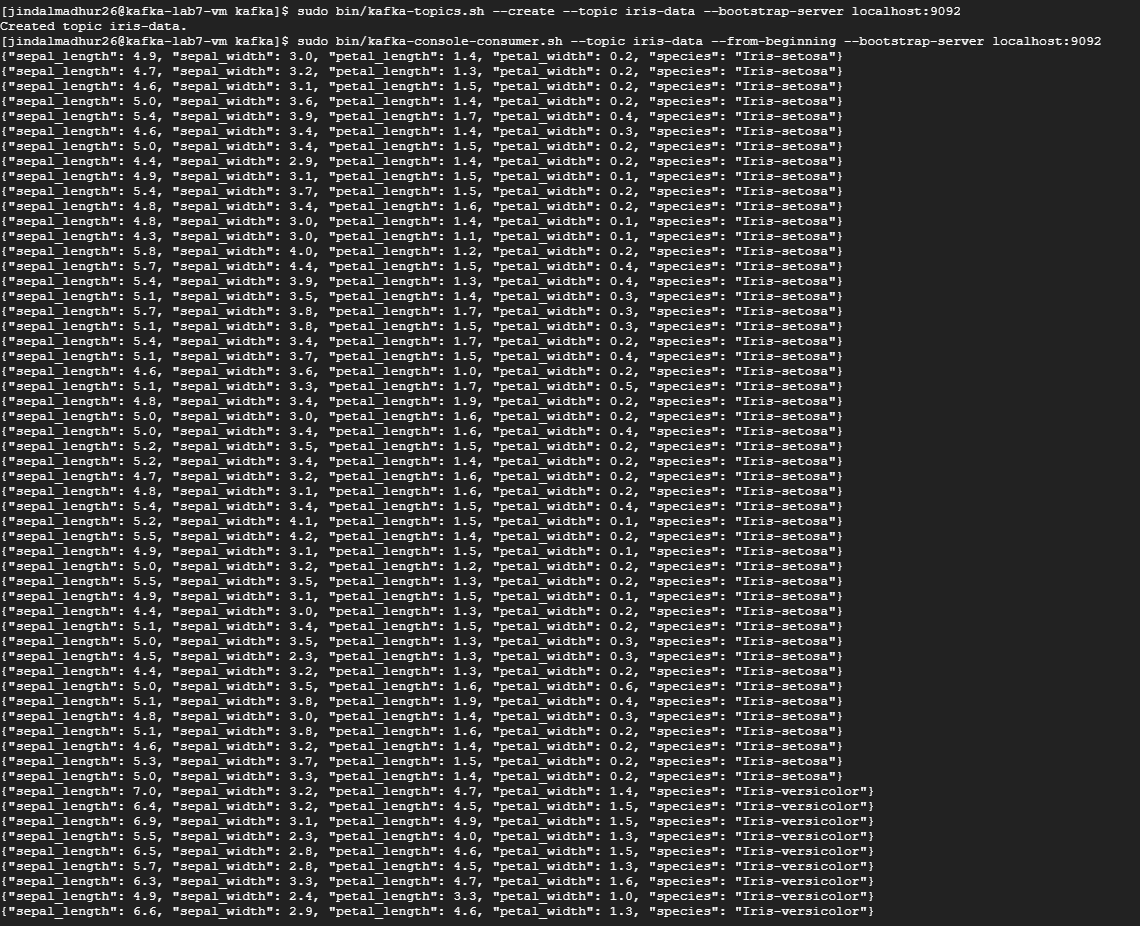
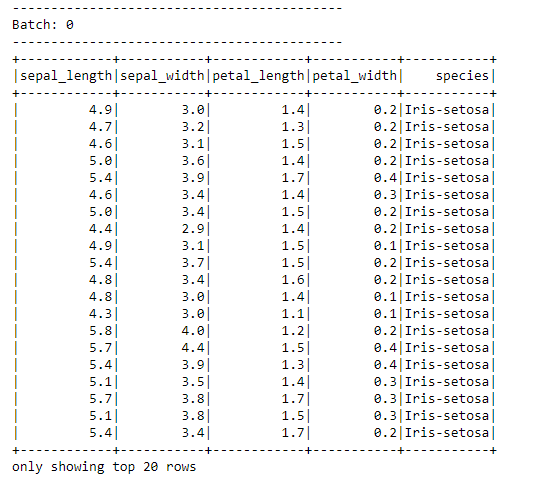
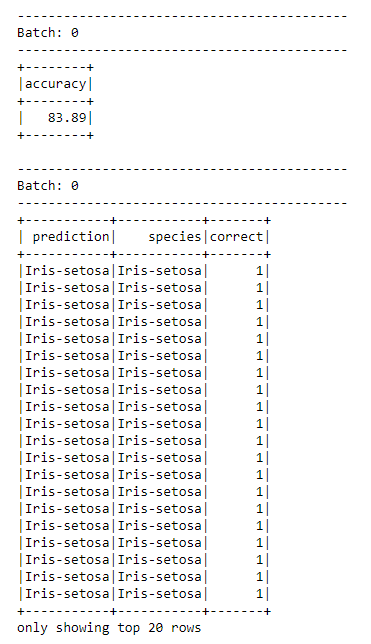
1. The producer\_json.py file contains the code for reading in the csv file and publishing messages as json format to the topic: iris-data



1. The subscriber\_json.ipynb contains the code for subscribing to the topic iris-data using Spark Structured streaming and connecting to Kafka service. The output for the dataframe created using the raw input reading from the topic is presented below.



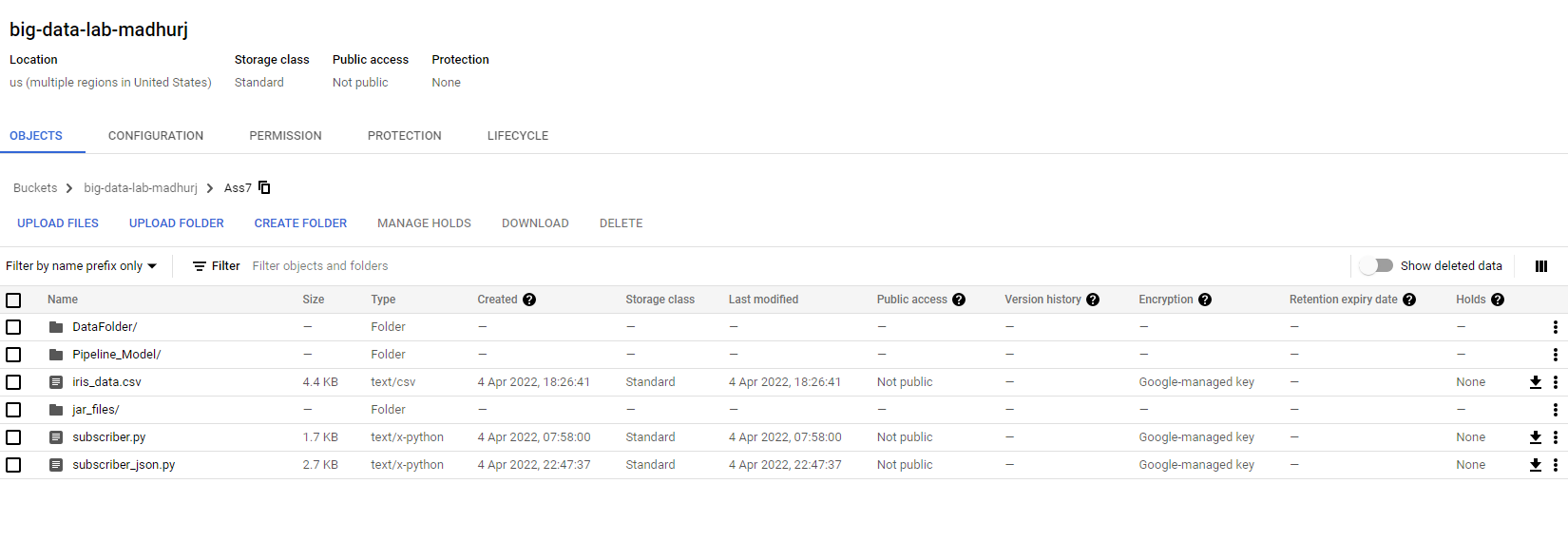
The dataframe is then transformed using the pipeline model, which gives us the prediction column which is used for calculating if the answer is correct (correct column) and then finally averaged to get the accuracy value. The screenshot depicts the above. A pdf file subscriber.pdf is attached which contains the pdf conversion of the ipynb file with the output.



P.s. this is performed in the jupyter notebook of the cluster because there was an error while submitting a dataproc pyspark job which is given below. This could be due to faulty jar file, not satisfying the dependencies which has been taken care of in the jupyter notebook. Below is the screenshot for the same.



1. The code for creating the Pipeline model is present in the pipeline\_model.py



The screenshot for the bucket.