**Task 3**

**Objective:** A simple ML project using spark ml library

Please refer to the data and README.txt files given in task3 folder

Prediction task is to determine whether a person makes over 50K a year or NOT.

* You have to prepare a script, hence please save all the commands from terminal at a safe place
* You may use any ML algorithm of your choice
* Please measure the accuracy using test.csv file
* The structure of a spark script code is given in the task3 folder. python\_code.py file

How to deploy spark code in production environment

**Code:**

if \_\_name\_\_ == "\_\_main\_\_":

#Initialize the spark session

spark = SparkSession \

.builder \

.appName("pyspark script job") \

.getOrCreate()

#Do all ETL here

df1 = spark.read.csv("/home/s\_kante/spark/data/agent\_system.csv", header=True)

df1.show()

#Once done, stop the spark session

spark.stop()

**How to execute:**

spark-submit <file name>.py