HOW TO PERFORM YOUR FIRST ETL ON SPARK CLUSTER

Step By Step guide with Planning on Jupyter

WHAT WE NEED?

1) THE COLUMN NAMES ARE MODIFIED IN THE SPARK DATAFRAME

2) NEW TABLE UNDER THE NAME CUSTOMER_SPARK_TABLE IS CREATED IN SPARK METASTORE

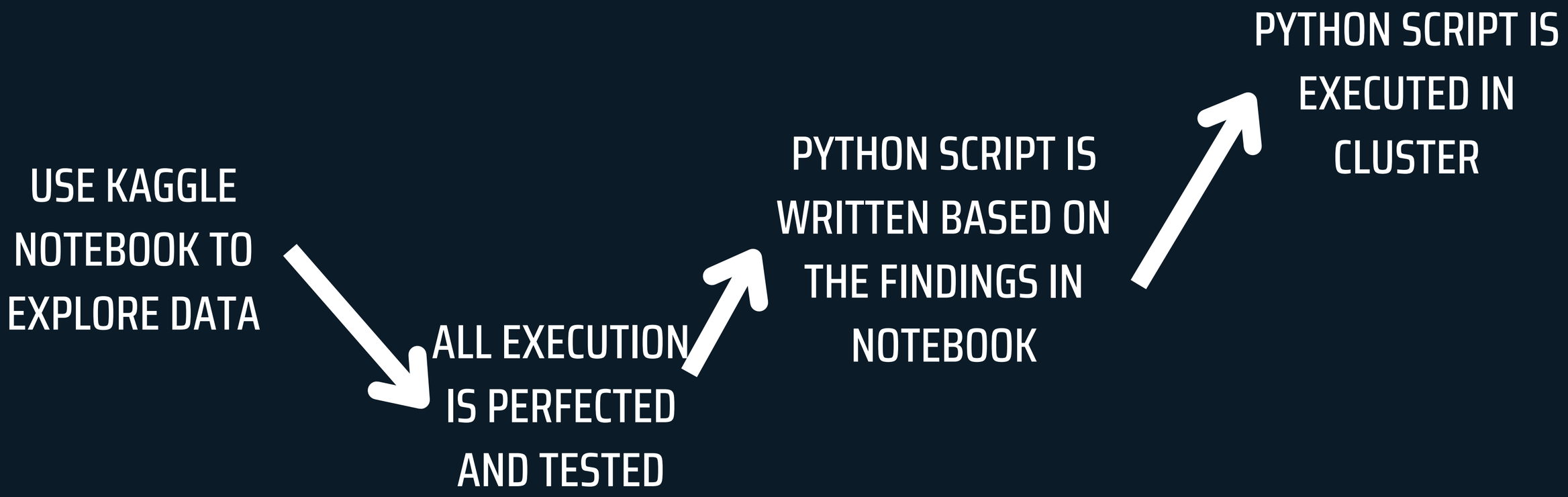
3) EXECUTE A SIMPLE FILTER TRANSFORMATION. SELECT THE ROWS THAT HAVE INCOME ABOVE 15000, AND SPENDING POWER ABOVE 50

4) WRITE A NEW TABLE INSIDE SPARK METASTORE

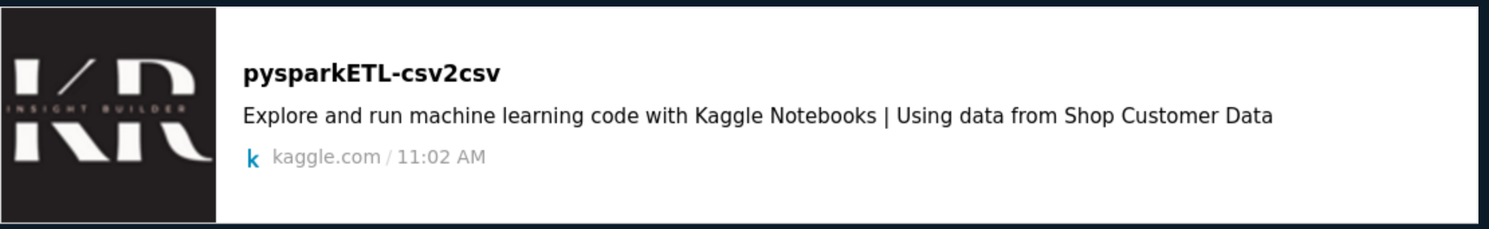
5) WRITE THE NEW TABLE AS CSV FILE

6) CONVERT THE JUPYTER NOTEBOOK CELLS INTO PYSPARK SCRIPT THAT CAN EXECUTE CODE ON THE GIVEN CSV FILE(IT WILL CUSTOMER.CSV FILE ONLY)

HOW WE ARE DOING IT?



LETS GET OURSELF A CLUSTER AND DIG IN



/OPT/SPARK3/SBIN/START-MASTER.SH

SPARK://IP-MASTER:7077

/OPT/SPARK3/BIN/SPARK-CLASS ORG.APACHE.SPARK.DEPLOY.WORKER.WORKER SPARK://IP-MASTER:7077

SPARK-SUBMIT --MASTER SPARK://IP-MASTER:7077 --CONF SPARK.SQL.WAREHOUSE.DIR=/USER/UBUNTU/EMR_INSTANCE_PRACTICE/PIPELINE_SCRIPTS/CUSTOMER_CSV2CSV.PY

THANKS FOR WATCHING

PRACTICE

PRACTICE

LIKE

SHARE

SUBSCRIBE

PRACTICE

PRACTICE

HTTPS://GITHUB.COM/INSIGHTBUILDER