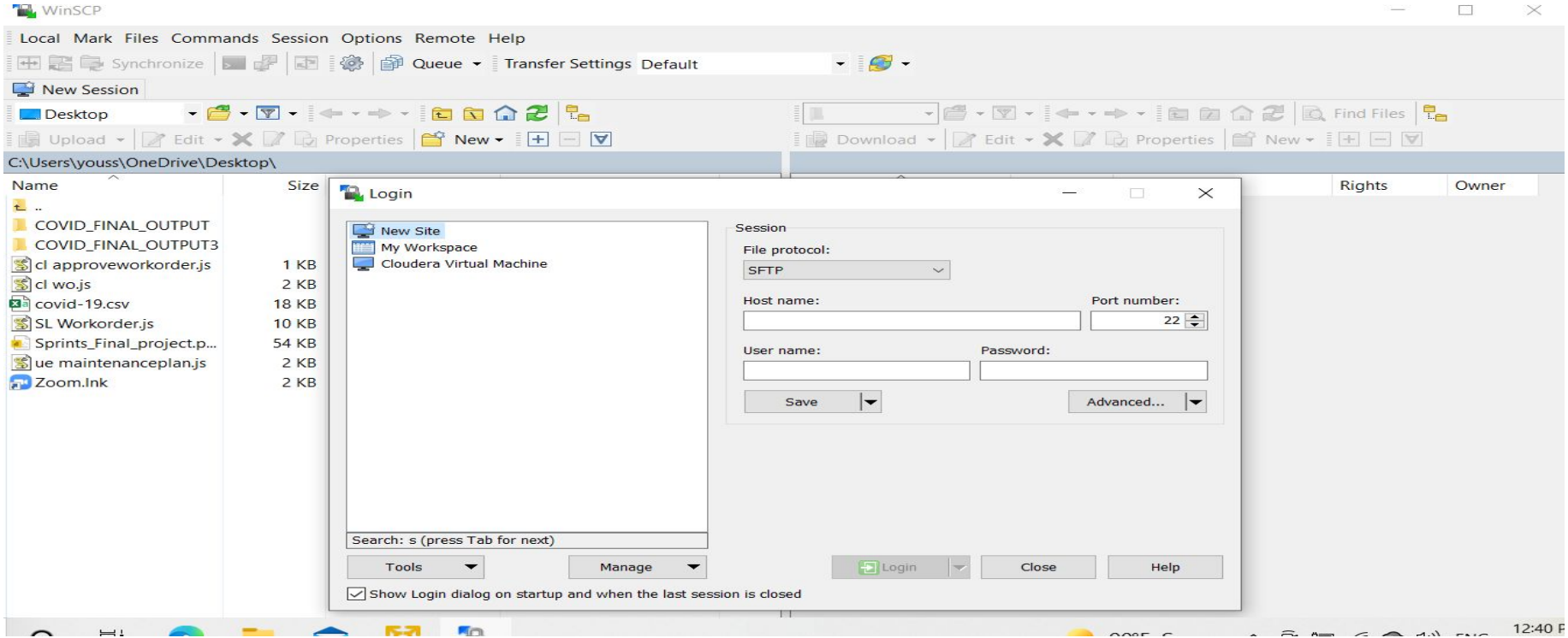
BIG DATA PROJECT

FIRST I STARTED BY UPLOADING THE CSV FILE TO THE MACHINE USING WINSCP



THEN I MOVED THE CSV FILE FROM THE VIRTUAL MACHINE FILE SYSTEM TO THE HADOOP FILE SYSTEM USING THIS COMMANDS

```
#!/bin/bash

#Landing Zones in Linux and HDFS
LINUX_LANDING_AREA=/home/cloudera/covid_project/landing_zone/COVID_SRC_LZ
HDFS_LZ=/user/cloudera/ds/COVID_HDFS_LZ


echo "GLOBAL Variables= " $LINUX_LANDING_AREA ", " $HDFS_LZ


hdfs dfs -mkdir -p $HDFS_LZ
echo "COVID_HDFS_LZ CREATED sucessfully"


hdfs dfs -put $LINUX_LANDING_AREA/covid-19.csv $HDFS_LZ
echo "covid-19.csv dataset LOADED sucessfully"
```

AFTER THAT I OPENED THE CLOUDERA THROUGH WEBSITE AND THE HIVE EDITOR I STARTED RUNNING THIS COMMANDS ONE AFTER ANOTHER

```
CREATE database covid_db;

use covid_db;

SET hive.exec.max.dynamic.partitions=500000;
SET hive.exec.max.dynamic.partitions.pernode = 500000;
SET hive.exec.dynamic.partition=true;
SET hive.exec.dynamic.partition.mode=nonstrict;

CREATE TABLE IF NOT EXISTS covid_db.covid_staging
(
 Country                                STRING,
 Total_Cases                                DOUBLE,
 New_Cases                              DOUBLE,
 Total_Deaths                           DOUBLE,
 New_Deaths                             DOUBLE,
 Total_Recovered                        DOUBLE,
 Active_Cases                           DOUBLE,
 Serious                               DOUBLE,
 Tot_Cases                             DOUBLE,
 Deaths                                DOUBLE,
 Total_Tests                                DOUBLE,
 Tests                                      DOUBLE,
 CASES_per_Test                         DOUBLE,
 Death_in_Closed_Cases                    STRING,
 Rank_by_Testing_rate                     DOUBLE,
 Rank_by_Death_rate                       DOUBLE,
 Rank_by_Cases_rate                       DOUBLE,
 Rank_by_Death_of_Closed_Cases      DOUBLE
)
ROW FORMAT DELIMITED FIELDS TERMINATED by ','
STORED as TEXTFILE
LOCATION '/user/cloudera/ds/COVID_HDFS_LZ'
tblproperties ("skip.header.line.count"="1");
```

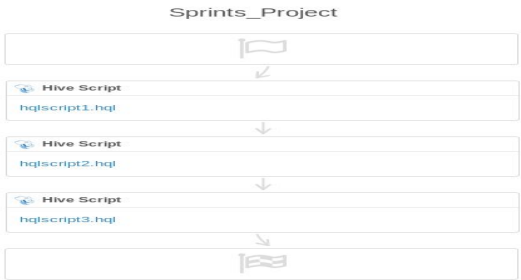
```
CREATE EXTERNAL TABLE IF NOT EXISTS covid_db.covid_ds_partitioned
(
 Country                                STRING,
 Total_Cases                                DOUBLE,
 New_Cases                              DOUBLE,
 Total_Deaths                           DOUBLE,
 New_Deaths                             DOUBLE,
 Total_Recovered                        DOUBLE,
 Active_Cases                           DOUBLE,
 Serious                               DOUBLE,
 Tot_Cases                             DOUBLE,
 Deaths                                DOUBLE,
 Total_Tests                                DOUBLE,
 Tests                                      DOUBLE,
 CASES_per_Test                         DOUBLE,
 Death_in_Closed_Cases                    STRING,
 Rank_by_Testing_rate                     DOUBLE,
 Rank_by_Death_rate                       DOUBLE,
 Rank_by_Cases_rate                       DOUBLE,
 Rank_by_Death_of_Closed_Cases      DOUBLE
)
PARTITIONED BY (COUNTRY_NAME STRING)
LOCATION '/user/cloudera/ds/COVID_HDFS_PARTITIONED';

FROM
covid_db.covid_staging
INSERT INTO TABLE covid_db.covid_ds_partitioned PARTITION(COUNTRY_NAME
SELECT *,Country WHERE Country is not null;

 DROP TABLE IF EXISTS covid_final_output;
CREATE EXTERNAL TABLE covid_db.covid_final_output
(
 Country                            STRING,
 TOP_DEATH                              STRING,
 TOP_TEST                           STRING
)
PARTITIONED BY (COUNTRY_NAME STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED by ','
STORED as TEXTFILE
LOCATION '/user/cloudera/ds/COVID_FINAL_OUTPUT';

FROM
covid_db.covid_ds_partitioned
INSERT INTO TABLE covid_db.covid_final_output PARTITION(COUNTRY_NAME)
SELECT Country,Rank_by_Testing_rate,Rank_by_Death_rate,Country
WHERE  Country is not null;
```

After working with hive editor I worked on oozie workflow I divided the full hive script into 3 scripts for creation of table ,staging and loading data into the final table named as hqlscript1 then 2 then 3 .



AFTER GETTING THE COVID FINAL OUTPUT FILE AND MOVING IT FROM THE HADOOP FILE SYSTEM TO THE VIRTUAL MACHINE FILE SYSTEM USING COMMAND COPYTOLOCALE AND THEN TO MY MACHINE I STARTED DOING THE VISUALIZATIONS