ONLINE BANKING ANALYSIS

This is the first project where we worked on apache spark, In this project what we have done is that we downloaded the datasets from KAGGLE where everyone is aware of, we have downloaded loan, customers credit card and transactions datasets . After downloading the datsaets we have cleaned the data . Then after by using new tools and technologies like spark, HDFS, Hive and many more we have executed new use cases on the datasets, that we have downloaded from kaggle. As we all know apache spark is a framework that can quickly process the large datsets.

So now let me explain the dataflow of how we have done is, first primarly we have ingested the data that is , we retrieved the data and then downloaded the datasets from kaggle and then we stored this datasets in cloud storage and imported from MYSQL to hive by sqoop this is how we have ingested the data , second after ingesting the data we have processed the large datasets in hive and then we have analyzed the data using pyspark in jupyter notebook by implementing several use cases.

In loandata.csv file
#number of loans in each category
#number of people who have taken more than 1 lack loan
#number of people with income greater than 60000 rupees
#number of people with 2 or more returned cheques and income less than 50000
#number of people with 2 or more returned cheques and are single
#number of people with expenditure over 50000 a month
#number of members who are elgible for credit card

In credit.csv file
#credit card users in Spain
#number of members who are elgible and active in the bank

In Transactions file
#Maximum withdrawal amount in transactions
MINIMUM WITHDRAWAL AMOUNT OF AN ACCOUNT in txn.csv
#MAXIMUM DEPOSIT AMOUNT OF AN ACCOUNT
#MINIMUM DEPOSIT AMOUNT OF AN ACCOUNT
#sum of balance in every bank account
#Number of transaction on each date
#List of customers with withdrawal amount more than 1 lakh