Market Analysis in Banking Domain

Background and Objective:

Your client, a Portuguese banking institution, ran a marketing campaign to convince potential customers to invest in a bank term deposit scheme.
The marketing campaigns were based on phone calls. Often, the same customer was contacted more than once through phone, in order to assess if they would want to subscribe to the bank term deposit or not. You have to perform the marketing analysis of the data generated by this campaign.

Domain: Banking (Market Analysis)

Dataset Description

The data fields are as follows:

1.	age	numeric
2.	job	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3.	marital	marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4.	education	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5.	default	has credit in default? (categorical: 'no', 'yes', 'unknown')
6.	housing:	has housing loan? (categorical: 'no', 'yes', 'unknown')
7.	loan	has a personal loan? (categorical: 'no', 'yes', 'unknown')

# related to the last contact of the current campaign:

8.	contact	contact communication type (categorical: 'cellular', 'telephone')
9.	month	Month of last contact (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10.	day_of_week	last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11.	duration	last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (example, if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call "y" is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:

12. campaign number of times a customer was contacted during the campaign (numeric, includes last contact)

13. pdays: number of days passed after the customer was last contacted from a previous campaign (numeric; 999 means customer was not previously contacted)

14. previous number of times the customer was contacted prior to (or before) this campaign (numeric)

15. poutcome outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

#Output variable (desired target):

16 y has the customer subscribed a term deposit? (binary: 'yes', 'no')

Analysis tasks to be done-:

The data size is huge and the marketing team has asked you to perform the below analysis-

1. Load data into HDFS, create Hive table
2. Create a Spark data frame by calling hive table
3. Perform below analysis using Hive & Spark.
    1. Give marketing success rate (No. of people subscribed / total no. of entries)
    2. Give marketing failure rate
        - Give the maximum, mean, and minimum age of the average targeted customer
        - Check the quality of customers by checking average balance, median balance of customers
        - Check if age matters in marketing subscription for deposit
        - Check if marital status mattered for a subscription to deposit
        - Check if age and marital status together mattered for a subscription to deposit scheme
        - Do feature engineering for the bank and find the right age effect on the campaign.

Note: Create end to end data pipeline using Bash Commands