Marketing Analysis

A Marketing campaign was run. It was based on phone calls. Often, the same customer was contacted more than ones through phone, in order to assess if they would subscribe to the bank term deposit or not.

The data fields are :

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 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')

Topics Covered:

Distributed Systems, Ingestion, Transformation/Processing, Data Warehousing, Stream Processing(optional), Visualization, Scheduling

Techstacks used:

S3/HDFS, PySpark/Spark, AWS Glue/ETL Tool of choice, Hive, SQL, Kafka(optional), Sqoop, Tableau/Power BI, Apache Airflow/Oozie

Task:

1. Load data and create Spark data frame
2.a) Give marketing success rate. (No. of people subscribed / total no. of entries) (Spark SQL)
   b) Give marketing failure rate
3. Maximum, Mean, and Minimum age of average targeted customer
4. Check quality of customers by checking average balance, median balance of customers
5. Check if age matters in marketing subscription for deposit
6. Check if marital status mattered for subscription to deposit.
7. Check if age and marital status together mattered for subscription to deposit scheme
8. Do feature engineering for column—age and find right age effect on campaign

9.Plot the ‘age’ vs ‘y’ using any visualisation tool

Steps to Perform:

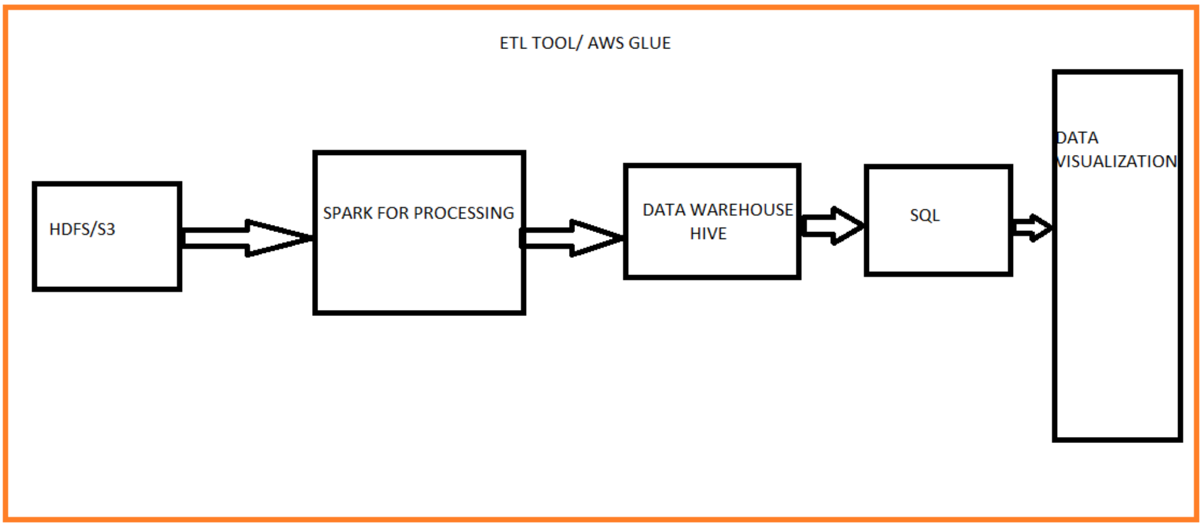
Though the data is stored in csv file but it is not properly formatted.

For performing the analysis, the data has to be pre-processed and then stored.

1) Store the data in S3 bucket/ HDFS.
2) Read the data from the storage space using Spark to perform the processing.
3) Perform the tasks mentioned above and move the processed data to Hive using proper partition and bucketing mechanism.
4) Move the data from hive to Sql database using sqoop so that it can be used for visualisation.
5) Now automate the entire pipeline using any ETL tool i.e AWS Glue or any tool.
6) Use a scheduler(Apache Airflow, Oozie) to run the scripts based on any specific window.

And if you want to create a streaming application, then use Kafka as a source and pull the data from Kafka topics to perform the processing of the data.

Proposed Pipeline for the project


ETL TOOL/ AWS GLUE
HDFS/S3
SPARK FOR PROCESSING
DATA WAREHOUSE
HIVE
SQL
DATA
VISUALIZATION