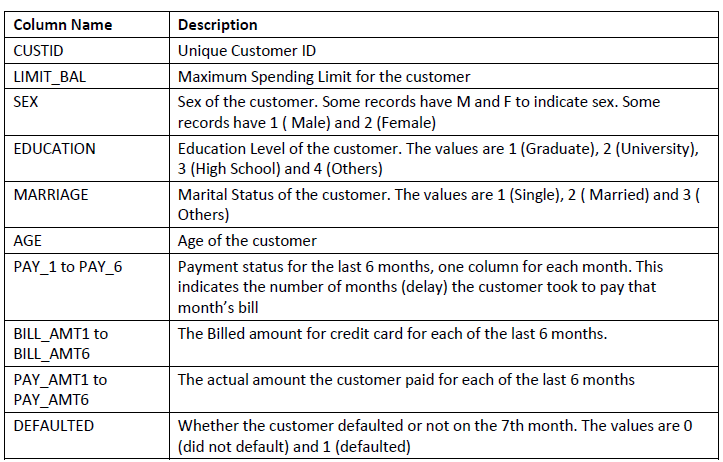
**Problem Statement**

This project corresponds to a real life scenario found in Credit Card companies. Customers who own credit cards are expected to pay off their balances monthly. But, they do default (not pay), which forces the bank into financial situations. Banks want to know which customer would possibly default in the future, so they can take necessary actions (such as closing their card, reducing their spending limits etc.). This problem involves a specific bank who wants to analyze their customer’s payment patterns and narrow down to cases where they are most likely to default.

This problem has a dataset that contains information about Credit Card customers for the past 6 and a set of questions that the bank has. Your assignment is to analyze the data and come up with answers to these questions using Apache Spark



**PR#01**: Do Data cleansing and enhancements as required to solve the problem. The dataset does have problems that you need to find out and fix. If you proceed without, you are going to see processing errors.

**PR#02**: Is there a clear distinction between Males and females when it comes to the pattern of defaulting? Do one sex default more than the other? Produce a report that looks like this showing percent defaulted for both males and females.

+--------+-----+--------+-----------+

|SEX\_NAME|Total|Defaults|PER\_DEFAULT|

+--------+-----+--------+-----------+

| Female| 591| 218.0| 37.0|

| Male| 409| 185.0| 45.0|

+--------+-----+--------+-----------+

**PR#03:** How does marital status and level of education affect the level of defaulting? Does one category of customers default more than the other? Produce a report that looks like the following.

+---------+-----------+-----+--------+-----------+

|MARR\_DESC| ED\_STR|Total|Defaults|PER\_DEFAULT|

+---------+-----------+-----+--------+-----------+

| Married| Graduate| 268| 69.0| 26.0|

| Married|High School| 55| 24.0| 44.0|

| Married| Others| 4| 2.0| 50.0|

| Married| University| 243| 65.0| 27.0|

| Others| Graduate| 4| 4.0| 100.0|

| Others|High School| 8| 6.0| 75.0|

| Others| University| 7| 3.0| 43.0|

| Single| Graduate| 123| 71.0| 58.0|

| Single|High School| 87| 52.0| 60.0|

| Single| Others| 3| 2.0| 67.0|

| Single| University| 198| 105.0| 53.0|

+---------+-----------+-----+--------+-----------+

**PR#04:** Does the average payment delay for the previous 6 months provide any indication for the customer to default in the future? Produce a report that looks like the following.

+-----------+-----+--------+-----------+

|AVG\_PAY\_DUR|Total|Defaults|PER\_DEFAULT|

+-----------+-----+--------+-----------+

| 0.0| 356| 141.0| 40.0|

| 1.0| 552| 218.0| 39.0|

| 2.0| 85| 41.0| 48.0|

| 3.0| 4| 2.0| 50.0|

| 4.0| 3| 1.0| 33.0|

+-----------+-----+--------+-----------+

**PR#05**: Come up with a prediction model that can predict whether the customer is going to default in the next month based on his/her history for the previous 6 months. Choose the best algorithm for this prediction model.

**PR#06:** The bank intends to group its customers into 4 groups based on their following attributes: SEX, EDUCATION, MARRIAGE, AGE\_RANGE (ranges of 10). Come up with an algorithm to do this grouping based on their affinity to each other.

Solution Hints

1. **Data Cleansing and Augmentation**

In this part, you will clean and prepare the data for further analysis

* Load the csv file into a data frame
* Remove the header lines
* The CSV file has junk characters in some rows. Remove them
* The CSV file has double quotes around certain values. Remove them.
* Write a conversion function that would convert this text RDD into a Row RDD of transformed data. Perform the following changes / transformations for the data
  + Create a new age variable where the age is rounded off to 10s. The age would be 10, 20, 30 etc. Required for PR#06
  + The Sex column contains both numeric (1, 2) and text representations (M, F). Normalize them to 1 and 2.
  + Compute average Billed amount (optional). These are things you try out additionally.
  + Compute average Pay amount (optional)
  + Compute average Pay duration. Make sure the values are positive. The dataset has a lot of negative values. This is required for PR#04
  + Compute Average Percentage paid as (average billed amount / average paid amount). This is to pursue a hypothesis that there is a possibility that this value might be able to predict defaulters. A low percentage paid “may” resulting in high defaulting. This is where you get creative with the solution. Feel free to try other ones too.
* Add a new column SEXNAME that contains Male and Female as values. Create a Data frame with those IDs and values and then join them with the main data frame. Required for PR#02
* Add a new column ED\_STR that contains an actual string for education. Create a Data frame with those IDs and values and then join them with the main data frame. Required for PR#03
* Add a new column MARR\_DESC that contains a description for marital status. Create a Data frame with those IDs and values and then join them with the main data frame. Required for PR#04

1. **Perform Analysis**

* Load the Data frame as a temp table /view
* Query the temp table to solve PR#02
* Query the temp table to solve PR#03
* Query the temp table to solve PR#04
* Perform correlation analysis

1. **Predict Defaulters ( PR#05 )**

* Prepare the data in the standard manner for machine learning

a. Convert to labeled point

b. Add indexing

* Split into training and test data sets.
* Run classification using 3 algorithms – namely Decision trees, Random Forests and Naïve Bayes. Find out which one gives the most accuracy on the test dataset.

1. **Group Data based on Attributes ( PR#06 )**

* Create a filtered dataset with only the attributes required for grouping.
* Perform centering and scaling on all the values
* Use KMeans clustering to group the data into 4 clusters.