**Description of the ccFraud dataset.**

**About Dataset**

The ccFraud dataset has a high volume which cannot be processed on a single machine. Thus compelling to the use of distributed processing using Apache Spark. This dataset is a snapshot at a particular instant of time for processing, as there is non-availability of credit card dataset having real time inflow of transactions in the public domain. The ccFraud dataset is a highly unbalanced dataset with only 5.98% of fraudulent transactions, rendering the veracity in the data. Making the variety and veracity present by the inclusion of portfolio data is an identified problem area.

The ccFraud dataset contains 1 million samples. We considered the genuine transactions as negative samples and fraudulent transactions as positive samples. The negative class has 985836 samples. The positive class has 62739 samples.

The dataset contains 9 features with a total size of 291.7Mb. The different variables in the dataset are:

(i) custID : customer ID, auto-incrementing integer value,

(ii) gender : taking two values either 1 or 2 for male and female,

(iii) state : state number given as integer,

(iv) cardholder: number of cards per customer with a maximum value of 2,

(v) balance : credit balance,

(vi) numTrans : number of transactions made in integer,

(vii) numIntlTrans : number of international transactions made in integer,

(viii) creditLine : credit limit of a customer in integer, and

(ix) fraudRisk : whether a given transaction is fraud or not.

The “fraudRisk” is a binary feature having 1 and 0 as two discrete values. Here, 1 represents a fraudulent transaction and 0 is used for a non-fraudulent transaction. And We discarded the “custID”, since it contains unique values in all samples, which will disturb in the generalization of patterns.