**ETL Project Report**

created by : Ritesh Vaish ( no partner)

The ETL project work has been divided into following activities:

**Extraction :**

* During the identification of project topic went through various sources of data and finally took 2 datasets from [Kaggle](<https://www.kaggle.com/>)
* Selected the two data files had the Data on Google stores Apps and the user sentiments and reviews
* As the data in both the files was in csv format , used pandas to extract the data from the csv files into panda’s data frame

**Transformation :**

* Examined the date and found that it needed cleaning
* Following cleanup was done on the datasets
  + Remove the records with had duplicates
  + Updated the null column’s with NaN
  + Removed all the records where any of the filed was NaN
* Following binning exercise was done on range of values in
  + ‘Rating’
  + ‘Price Range’
  + Sentiment polarity ,
  + Sentiment Subjectivity,
  + Changed the data to utf8 characterset

**Loading :**

* Once the date was cleaned and added binning details , renamed the dataframe colums to lowercase, renamed some columns which had spaces and names in upper case
* Created a new DB and tables with corresponding tables names in it
* Using pandas created connection to the mysql database on local host
* Loaded the tables in mysql and later extracted the data from tables for validating the data.