Using Hive :

Step 1 : creating a Database called assignment1

Command : create database assignment1;

Use assignment1 ; – to use the database for our following task

Step 2 : creating table for the dataset

2.1 : Creating a table called “books\_rating”

Command :

create table books\_rating

( Id VARCHAR(100), Title STRING, Price DECIMAL, User\_id VARCHAR(30), profileName VARCHAR(200),review\_helpfulness VARCHAR(20), review\_score DECIMAL, review\_time DATE, review\_summary VARCHAR(50), review\_text VARCHAR(1000) )

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"separatorChar" = ",",

"quoteChar" = "\"")

tblproperties("skip.header.line.count"="1");

2.2 : Loading the data into table :

Command :

LOAD DATA LOCAL INPATH

'/home/aniket\_jagtap2/dataset/Books\_rating.csv'

INTO TABLE books\_rating;

Step 3: Cleaning

3.1 : Remove Duplicate Rows from the dataset.

INSERT OVERWRITE TABLE books\_rating

SELECT books\_rating.Id, books\_rating.Title , books\_rating.Price, books\_rating.User\_id, books\_rating.profileName , books\_rating.review\_helpfulness , books\_rating.review\_score , books\_rating.review\_time , books\_rating.review\_summary , books\_rating.review\_text

FROM books\_rating

GROUP BY books\_rating.Id, books\_rating.Title , books\_rating.Price, books\_rating.User\_id, books\_rating.profileName , books\_rating.review\_helpfulness , books\_rating.review\_score , books\_rating.review\_time , books\_rating.review\_summary ,

books\_rating.review\_text;

**Task 4.1: Query processed data to differentiate ham and spam part of the dataset**

Step 1:

Created UDF python code to detect spam using bag of words(for spam) :

**Spam\_detection.py -** python file name

@outputSchema(“num:int”)

Def total\_count(review\_text):

Total\_cnt = 0

bag = [...some bag of spam words…] #as a list

for i in bag:

for x in review\_txt.split():

If i==x:

total\_cnt = total\_cnt + 1

return total\_cnt

About the code : it iterates through all the words from a particular review(review\_text) and compare with the bag of spam words. if a word matches, the cnt variable gets incremented and it returns the total number of matches (in other words, number of spam words in a review)

Step 2 : Registered the python code using jython in PIG

REGISTER 'spam\_detection.py' USING jython AS spam\_finder;

loading the cleaned data into the pig

reviews = LOAD '/user/assignment1data' using PigStorage(',') as (id:int, title:chararray, price:float, user\_id:chararray, profileName:chararray, review\_helpfulness:chararray, review\_score:float, review\_time:long, review\_summary:chararray, review\_text:chararray);

Find Spam

spam = FILTER reviews BY ((id==’’) AND (profileName==’’)) ;

STORE spam\_reviews INTO ‘/user/assignment1data/spam\_data’;

Find ham

ham = FILTER reviews by 2<spam\_detection.py.total\_count(wordsbag.BOW.reviews.review\_text);

STORE ham INTO ‘/user/assignment1data/ham\_data’;

wordsbag = LOAD ‘/home/aniket\_jagtap2/bagofwords.txt’ as boww;