# 

# Assignment - I

# SPAM DETECTION SYSTEM

# CA675 - Cloud Technologies

# Yakoob Youseph Cyrus

# MCM1 2022-23

# Student No : 22263741

# [yakoob.yousephcyrus2@mail.dcu.ie](mailto:yakoob.yousephcyrus2@mail.dcu.ie)

# Link to github Repo: https://github.com/cyromoto/hadoop-01-cloud-tech

# Link to Cloud Project:- https://console.cloud.google.com/home/dashboard?project=hadoop-assignment-101

# Task 1 (Cloud Infra Setup)

# Installation of Hadoop and creation of cluster

# The cluster creation is done the Google Cloud Platforms (GCP) Compute Engine feature. A new compute engine named ‘hadoop-manual’ is created first for performing this task. An Ubuntu 18.04 OS was mounted on the instance since I found that the Ubuntu had a better community support when working with Hadoop. GCP offers a command line interface , attached below

# *gcloud compute instances create hadoop-instance --project=hadoop-assignment-101 --zone=us-central1-a --machine-type=e2-medium --network-interface=network-tier=PREMIUM,subnet=default --maintenance-policy=MIGRATE --provisioning-model=STANDARD --service-account=159560514124-compute@developer.gserviceaccount.com --scopes=https://www.googleapis.com/auth/cloud-platform --create-disk=auto-delete=yes,boot=yes,device-name=hadoop-instance,image=projects/ubuntu-os-cloud/global/images/ubuntu-1804-bionic-v20221018,mode=rw,size=10,type=projects/hadoop-assignment-101/zones/us-central1-a/diskTypes/pd-balanced --no-shielded-secure-boot --shielded-vtpm --shielded-integrity-monitoring --reservation-affinity=any*

# Created the Compute Engine using the GUI.

# After creation of Instance, Java and Hadoop were installed.

# Necessary PATH variable changes were made in *.bashrc* file

# A Pseudo distributed mode was set up in the instance

# For the set up of pseudo distributed mode necessary changes in different confifuration files were made,this includes changes in core-site.xml,hdfs-site.xml,yarn-site.xml,mapred-site.xml.

# Hadoop installation is verified by starting it with

$ start-dfs.sh

# The results of the queries are attached below

# 

# The above shows a cluster created with a namenode and a datanode in a pseudo distributed mode.

# Pig Installation: The latest version of pig was installed to the instance [Pig(0.17)](https://dlcdn.apache.org/pig/pig-0.17.0/)

# 

# Hive Installation :The latest version of Hive was installed . [Hive 3.1.3](https://dlcdn.apache.org/hive/hive-3.1.3/)

# 

# The Installations can be verified on the instance using the super user ‘hadoop’ with the password as ‘password’

# Task 2 (Dataset selection and setup)

# The dataset was selected from Kaggle. The dataset comprised of amazon reviews for the music’s Amazon offered. The dataset is publicly available and link to the dataset is given below:

# <https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset>

# The dataset was chosen since it had sufficient number of records and necessary attributes which could help in the classification as ham and spam namely the review body and the review subject. The dataset is a subset of a review dataset which contains reviews of different classes of items (phones, apparels, electronics etc).

# The dataset was uploaded to the google cloud storage bucket

# *bucket-01/amazon-music/amazon\_reviews\_us\_Music*

# A new directory named dataset was created In the hdfs and the above uploaded file was copied from the bucket to the folder

# 

# *hdfs dfs -cp gs://hadoop-data-bucket-01/amazon-music/amazon\_reviews\_ us \_Music.tsv /dataset*

# This file is used for further processing’s like cleaning and analysis.

# Task 3(Clean & Process Using HIVE/PIG)

# The data is cleaned using Apache Hive. First, table named *musicset* is created in the database from the tsv file in hdfs. Creation is done using the following query

# *CREATE EXTERNAL TABLE*

# After the creation of table the necessary cleaning is done on the table.

# Removal of unwanted characters and unnecessary columns.

# 

# The hive queries are uploaded in [GitHub](https://github.com/cyromoto/hadoop-01-cloud-tech/blob/main/CloudAssignment/_3/cleaning.hql).

# Task 4(HAM /SPAM using PIG/HIVE)

# For the getting a list of spam and ham a pre classified set of ham and spam is used. From this using a python script a bag of words are generated for each spam and ham.The link to the data set is given below:-

# <https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset>

# The review body of the reviews are matched for keywords from the bag of words created from the pre classified data set. The classified data is grouped using the customer ID’s . and ordered. From this the top are selected from each set (HAM and SPAM). The queries are written in PigLatin

# TOP 10 SPAM ACCOUNTS ID

# 

# TOP 10 HAM ACCOUNTS ID

# 

# Task 5(TF-IDF using MapReduce)

# The top spam words and top ham words are chosen from the respective spam and ham tables created. The [mapper](https://github.com/cyromoto/hadoop-01-cloud-tech/blob/main/CloudAssignment/_5/mapper.py) and [reducer](https://github.com/cyromoto/hadoop-01-cloud-tech/blob/main/CloudAssignment/_5/reducer.py) are written in python language and the links are provided. They were tested in the local machines and results were generated

# TOP SPAM WORD

# TOP HAM WORD

# The command to execute the streaming jar and provide the mapper and reducer are given below:-

# *hadoop jar hadoop-streaming-3.2.4.jar -file /home/yakoob\_yousephcyrus2/mapper.py -mapper "python mapper.py" -file /home/yakoob\_yousephcyrus2/reducer.py -reducer "python reducer.py" -input hdfs://10.128.0.23:8020/mergedHam/- -output hdfs://10.128.0.23:8020/hammer/output*

# But the execution failed due to a python error

# 

# 

# Later while going through the logs in Hadoop, it was seen, as above, that the error was due to a failed import. Tried importing in other methods but wasn’t fruitful.

# References

# <https://www.tutorialspoint.com/apache_pig/index.htm>

# <https://stackoverflow.com/questions/39555547/pig-equivalent-of-like>

# <https://stackabuse.com/count-number-of-word-occurrences-in-list-python/>

# <https://stackoverflow.com/questions/952914/how-do-i-make-a-flat-list-out-of-a-list-of-lists>

# <https://stackoverflow.com/questions/20461165/how-to-convert-index-of-a-pandas-dataframe-into-a-column>