Big Data 18CS322

Assignment 3

Analysis of QuickDraw images with Spark

Assignment Objectives and Outcomes

a. The objective of this assignment is for the students to install **Spark** in pseudo distributed mode and to become familiar with the Spark programming environment and the RDDs.

b. At the end of this assignment, the student will be able to write and debug **Spark** code.

Ethical practices

Please submit original code only. You can discuss your approach with your friends but you must write original code. All solutions must be submitted through the portal. We will perform a plagiarism check on the code.

The Dataset:

a. We shall provide a link to the dataset on PESU forums. The schema for the dataset is shared below

Shape.csv

Key	**Type**	**Description**
key_id	Numeric string	A unique identifier across all drawings.
word	string	Category the player was prompted to draw.
countrycode	string	A two letter country code of where the player was located.

Shape_stat.csv

Key	Type	Description
key_id	Numeric string	A unique identifier across all drawings.
word	string	Category the player was prompted to draw.
recognized	boolean	Whether the word was recognized by the game.
timestamp	datetime	When the drawing was created.
Total_Strokes	Numeric	Represents the number of strokes used for the drawing

b. Use the dataset for the tasks given below.

c. The dataset column ordering is subject to change. Please do not hardcode columns. For example, it is incorrect to assume key_id is column 0.

Software/Languages to be used:

a. Spark - 3.0.1

b. Hadoop - 3.2.1

c. Code can only be in Python

Marks:

a. Task 1 : 2 marks

b. Task 2 : 2 marks

c. Viva : 1 mark

Submission Date:

a. To be announced on the forum

Tasks Overview:

a. Join operation with and without co-partitioning
b. Observe the DAG for both the tasks
c. Submit one page report based on the template and answer the questions on the report.

Submission Link:

Will be shared with you through PESU Forums along with the report template.

Task Specifications:

a. Task 0:

i. Install **Spark** in Pseudo-Distributed mode.

ii. You can refer the following few links (but not limited to) to help you get started - (Make sure that you install the correct versions of any software - as mentioned above)

- https://medium.com/@msris108/how-to-setup-a-pseudo-distributed-cluster-with-hadoop-3-2-1-and-apache-spark-3-0-34406a85130f
- https://ashwin.cloud/blog/running-apache-spark-on-a-single-node-pseudo-distributed-hadoop-cluster-in-macos/
- http://why-not-learn-something.blogspot.com/2015/06/spark-installation-pseudo.html

iii. You might need to use StackOverflow extensively to complete this task.

iv. See the setup video shared..

b. Task 1:

i. Problem Statement:

Find average number of strokes used to draw any given object.

ii. Description:

- Find the number of strokes for a given word, that has been marked as recognized.
- Find the number of strokes for the same word, that was not marked as recognized.

iii. Comments:

- Make sure the given word is passed as a command line argument - do not hardcode it.

iv. Output Format

Each output for the subtasks should be a new line.

The command used to run this task would be

spark-submit <filename.py> <word> hdfs://localhost:9000/dataset1 hdfs://localhost:9000/dataset2

Eg. Assuming the word given on the command line is bat, 10 is the avg. Strokes for recognized bat and 7 is the avg. Strokes for unrecognized bat.

10
7

PLEASE NOTE: These are representative outputs. Actual outputs may differ.
Please ensure the datasets are read using system arguments in your code.

c. Task 2:

i. Problem Statement:

Find object count by country based on number of strokes and shape recognition.

ii. Description:

Given a specific word, count the number of occurrences of that word per country.

iii. Comments

- Details for the drawing/shape are available in two (csv) files.
- To get country-wise result, it is necessary to join the two files.

- **Perform the join**
  - **With given input files as it is.**
  - **Co-partition the input files before join.**
- Observe the change in the two cases of join, and compare the wall clock times you get.
  - This will not be tested, so please use the first approach only for the submission. Your report should reflect the wall clock time information and the justification for the same.
- Output should be sorted in alphabetical order of countries
- An occurrence of the word should only be considered if it is unrecognized and number of strokes for that object < k
- Both specified word and stroke count k should be command line arguments. The first argument should be the word, and the second argument should be the stroke count - k.

iv. Output format

- Comma separated values in the form of:
  Country_code, count_of_word

- Each pair must be in a new line, with no space between the values
  spark-submit <filename.py> <word> <strokes> hdfs://localhost:9000/dataset1 hdfs://localhost:9000/dataset2

  Eg. (The values below are just for representational purposes)
  AE,3
  AU,53
  BR,43
  NY,345