DSC 204A: Scalable Data Systems

Programming Assignment 1

Released: 26 April 2023, Due: 10 May 2023

1 Introduction

In this assignment, you will be using Dask library to explore task parallelism on multiple machines. You will be performing feature explorations, data consistency checks, and computing several descriptive statistics about the dataset.

2 Dataset Description

You are provided with the Amazon Reviews dataset with the *reviews* and *products* tables as CSV files. The schemas are provided in Table 1.

(A) Column name	Column description	Example
reviewerID	ID of the reviewer	A32DT10X9WS4D0
asin	ID of the product	B003VX9DJM
reviewerName	name of the reviewer	Slade
helpful	helpfulness rating of the review	[0, 0]
reviewText	text of the review	this was a gift for my friend who loves touch lamps.
overall	rating of the product	1
summary	summary of the review	broken piece
unixReviewTime	unix timestamp of review	1397174400
reviewTime	time of the review (raw)	04 11, 2014

(B) Column name	Column description	Example
asin	ID of the product	143561
salesRank	sales rank information	{'Movies & TV': 376041}
imUrl	url of the product image	http://g-ecx.images-amazon.com / 31mC.jpg
categories	list of categories the product	[['Movies & TV', 'Movies']]
title	name of the product	Everyday Italian (with Giada de Laurentiis)
description	description of the product	3Pack DVD set - Italian Classics
price	price in US dollars	12.99
related	related products (also bought, also viewed, bought together, buy after viewing)	{'also_viewed': ['B0036FO6SI', '000014357X'],'buy_after_viewing': ['B0036FO6SI', 'B000KL8ODE']}
brand	brand name	Big Dreams

Table 1: (A) *Reviews* table and (B) *Products* table

3 Tasks

You will compute several descriptive statistics for both *reviews* and *products* table as follows:

Q1. Get percentage of missing values for all columns in the *reviews* table.

Q2. Get percentage of missing values for all columns in the *products* table.

Q3. Find Pearson correlation coefficient between the price and rating of the products.

Q4. Find mean, standard deviation, median, min, and max for the price column in the *products* table.

Q5. Find number of products for each super-category (the first entry in the "categories" column in the products table). Output categories should be sorted in non-increasing order in the number of products.

Q6. Check (return 1 or 0) if there are any dangling references from product ids in the *reviews* table to *products* table. Return 1 if there are dangling references and 0 otherwise.

Q7. Check (return 1 or 0) if there are any dangling references from product ids in the "related" column to the "asin" column of the *product* table. Return 1 if there are dangling references and 0 otherwise.

A code stub with the function signature has been provided to you. The input to the function is the *reviews* CSV file and the *products* CSV file. The expected data types of the output answers for each question has also been given in the code through assertion statements. Your answers must match these expected output data types. The answers are later converted to json and written to a file. We will time the execution of the function `PA1` which is stored in the "runtime" variable for your reference.

We have shared with you the "development" dataset and our accuracy results. Our code's runtime on 1 node and 4 nodes are roughly 395s and 115s respectively. You can use this to validate your results and debug your code. The final evaluation will happen on separate held-out test sets. The runtime and the speedup numbers will be different for the held-out test set.

4 Deliverables

Submit your source code as `<NAME>_<PID>.py` on Canvas. Your source code must conform to the function signatures provided to you. Make sure that your code is writing results to `results_PA1.json`.

5 Hints

1. In the "starter_code", you will notice I create my `Client` object like `client = Client('127.0.0.1:8786')` and not just `client = Client()` like we did for PA0. This is because, I already started my Dask scheduler on port 8786 in another terminal using the `dask scheduler` command. By passing `'127.0.0.1:8786'` into `Client`, I am telling Dask to use this scheduler and not create a new one.

2. We advise you to first create a smaller subset of the data (∼5000 rows) and use 5 instances so that you can test your code faster. This aligns with the "fail-fast" philosophy. Once your code is working you can measure the runtime on single and five instances and make further optimizations.

3. What are dangling references?
If there is a value V in column X of table A, but this value is missing in column X of table B, we say that V is a dangling reference of X from A to B.

For example consider the below two tables.

ID	X
1	21
2	32
3	1

ID	Y
1	4
3	5

Here, ID *2* is a dangling reference from the first table to the second.

4. Many of the answers need to be a Python dictionary. Consider methods like `to_dict()` to convert a Pandas Series to a dictionary.

5. Worker logs will be output to the terminals where you started your workers and may not appear in jupyter-notebook. You will need to check the terminal output for your workers to debug any errors.