CS3.304 - AOS Assignment 5

Introduction to Py-SPARK

Questions

You have to perform the following tasks in House_pricing dataset (Click here to download)

> ***About dataset***
> The Price Paid Data includes information on all registered property sales in England and Wales that are sold for full market value.
> The main fields of the dataset which are needed for Assignment are ⇒ ‘Transaction unique identifier’, ‘Price’, ‘Country’

Question 1: Second Highest Value Transaction(’Price’) in Selected Countries
Among the countries ‘GREATER LONDON’, ‘CLEVELAND’, ‘ESSEX’ find the second highest property sold.

Ex:

Consider the countries ‘India’, ‘Bangladesh’ from below table.

Transaction	Price	Country
00001	50,000	India
00002	35,000	India
00003	47,500	India
00004	37,000	Bangladesh
00005	48,000	Bangladesh
00006	29,000	Bhutan

Then result should be “Second highest transacted value is 48,000 which is transacted in Bangladesh’.

Question 2: Country with the Second Most Transactions
Among all the countries, find the one which has second most transaction count.

Ex:

For above example

As Bangladesh has 2nd highest number of transactions, the following should be the output -

“Country with second most transactions is ‘Bangladesh’.”

Note: If there are multiple countries with same number of second highest Transactions, you can report any one of them.

Question 3: Number of Transactions for each country
List out the number of transactions for each country into a csv file.

Ex:

For above example, the resultant csv file should contain entries -

Country	Count
India	3
Bangladesh	2
Bhutan	1

Instructions on using Abacus

- Login to the system using the credentials provided to you.

  Ex:-

  ```
  ssh cs3304.01@abacus.iiit.ac.in
  <Password> upon prompt.
  ```

- *Create Virtual Environment and install pyspark over it.*

  `pip install pyspark`

- Copy the files to the Abacus system -

```
scp <source_file_path> <user_id>@abacus.iiit.ac.in:<destination_path>
```

For copying files from Abacus to local system just use similar command as above by interchanging Source-Target

- **Execution of python program**
  - Requesting for an interactive Job

    Firstly request a Job over which you can execute the python program - `sint3 -c <num_cores>`

    > 📢 Note:- The requested number of cores should be multiple of 2 (2, 4, 6).

  - Once the Job is created, execute your python program as required.
  - Then Relinquish the Job using `exit`

> 📢 NOTE:- Ensure to Relinquish/exit the assigned Job after usage.

Instructions for Questions

All the questions are to be done using pyspark library functionalities.

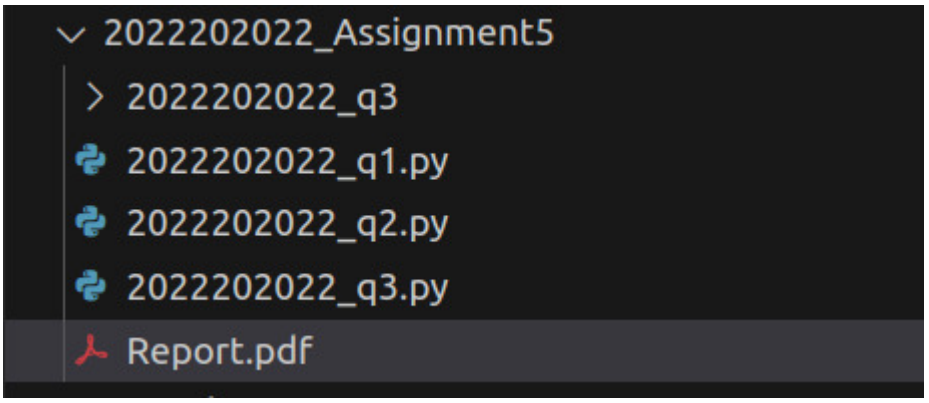
For each question, include the result into Report document.

For each question, you are expected to observe the **time taken by the program for execution by varying the number of cores to be used (2, 4 & 6)**. Finally report the comparison of time taken in the report for each question.

Also report your observations about the time comparison.

For Question 3, as there may be many rows for the output, you need to write the final result into a csv file and upload it along with the submission under a folder (<roll_number_q3>).

Submission Format



Note

- All the students are shared with login credentials (via mail) to use the abacus system.
- Report should have all the Results, Time comparisons and Clear explanation of your observations.
- There will not be any extension for the Assignment. Plan early to complete it by the deadline.