DSCI 525 - Web and Cloud Computing

Milestone 1: Tackling big data on your laptop

Overall project goal and data

During this course, you will be working on a team project involving big data. The purpose is to get exposure to working with much larger datasets than you have previously in MDS. You have been assigned to teams of three or four. (See group assignment in Canvas. Unlike previous project courses, in this course, all of you will be working on **the same problem**. In particular, you will be building and deploying ensemble machine learning models in the cloud to predict daily rainfall in Australia on a large dataset (~6 GB), where features are outputs of different climate models, and the target is the actual rainfall observation.

You will be using this dataset on figshare. This folder has the output of different climate models as features, and our ultimate goal is to build an ensemble model on these outputs and compare the results with the actual rainfall. At the end of the project, you should have your ML model deployed in the cloud for others to use.

During this course, you will work towards this goal step by step in four milestones.

Milestone 1 checklist

Part of the purpose of this milestone is to annoy you by making you work with large data in `Pandas` and vanilla CSV files. Typically these are not the best for dealing with large data. Along the way, you will also explore some useful tools for working with big data.

1. Team-work contract

rubric={correctness:10}

Similar to what you did in DSCI 522 and DSCI 524, create a teamwork contract. The contract should outline how you are committed to working together so that you are accountable to one another. Again, you may start with your team contract document from previous project courses and adapt it to your new team. It is a fairly personal document, and please do not push it into your public repositories. Instead, save it somewhere your team can easily share it, and you can share a link to it or a copy with us in your submission to Canvas to prove you did this.

2. Creating a repository and project structure

rubric={mechanics:10}

1. Similar to previous project courses, create a public repository under UBC-MDS org for your project.
2. Write a brief introduction of the project in the `README`.
3. Create a folder called `notebooks` in the repository and create a notebook for this milestone in that folder.

3. Downloading the data

rubric={correctness:10}

1. Download the data from figshare to your local computer using the figshare API (you need to make use of `requests` library).
2. Extract the zip file, again programmatically, similar to how we did it in class.

   You can download the data and unzip it manually. But we learned about APIs, so we can do it in a reproducible way with the `requests` library, similar to how we did it in class.

   There are 5 files in the figshare repo. The one we want is: `data.zip`

4. Combining data CSVs

rubric={correctness:10,reasoning:10}

1. Combine data CSVs into a single CSV using pandas.
2. When combining the CSV files, add an extra column called "model" that identifies the model. Tip 1: you can get this column populated from the file name, eg: for file name "SAM0-UNICON_daily_rainfall_NSW.csv", the model name is SAM0-UNICON Tip 2: Remember how we added "year" column when we combined airline CSVs. Here the regex will be to get word before an underscore ie, "/([^_]*)"

   Note: There is a file called `observed_daily_rainfall_SYD.csv` in the data folder that you downloaded. Make sure you exclude this file (programmatically or just take out that file from the folder) before you combine CSVs. We will use this file in our next milestone.

3. ***Compare*** run times on different machines within your team and summarize your observations.

Warning: Some of you might not be able to do it on your laptop. It's fine if you're unable to do it. Just make sure you discuss the reasons why you might not have been able to run this on your laptop.

5. Load the combined CSV to memory and perform a simple EDA

rubric={correctness:10,reasoning:10}

1. Investigate at least two of the following approaches to reduce memory usage while performing the EDA (e.g., value_counts). Refer to lecture notes here.
    - Changing `dtype` of your data
    - Load just columns that we want
    - Loading in chunks
2. ***Compare*** run times on different machines within your team and summarize your observations.

6. Perform a simple EDA in R

rubric={correctness:15,reasoning:10}

1. Choose one of the methods listed below for transferring the dataframe (i.e., the entire dataset) from Python to R, and explain why you opted for this approach instead of the others.
    - Parquet file
    - Pandas exchange
    - Arrow exchange
2. Once you have the dataframe in R, perform a simple EDA.

Specific expectations for this milestone

- In this milestone, we are looking for a well-documented and self-explanatory notebook that explores different options to tackle big data on your laptop.
- Please discuss any challenges or difficulties you faced when dealing with this large amount of data on your laptop. You can stop combining the data if it takes more than 30 minutes. Briefly explain your approach to overcoming the challenges or reasons why you could not overcome them.

- For questions 5 and 6, you are free to choose any exploratory data analysis (EDA) task you want. Visualization is not necessary; summarizing the data is enough. However, if you want to install additional packages for visualization that are not included in the .yml file, feel free to install them on top of your notebook. If you want to install packages in R, you can do so using `install.packages("dplyr")` under `%%R` magic cell.
- If someone in your team is facing issues with using R in a Python notebook, you can ignore it, as you will not need it for any other milestones. The main purpose of showing it in the lecture was to introduce and get a feel for the serialization and deserialization concept.
- You only need to ***compare*** the time with other team members for questions 4 and 5. You do not need to do this for question 6. You can use the following table to record your results. Feel free to add any other relevant columns.

Team Member	Operating System	RAM	Processor	Is SSD	Time taken
Member 1					
Member 2					
Member 3					
Member 4					

Submission instructions rubric={mechanics:5}

In the textbox provided on Canvas for the Milestone 1 assignment include:

- The GitHub URL to your notebook.

As comment include - Repo link - Teamwork contract