Ass 02: Election 2024 Analyzer

Task 0: Dataset Analysis

You will be downloading the election2024.csv file containing the following columns:

Type:	National	Punjab
	Khyber Pakhtunkhwa	Sindh
	(Balochistan data not available, but if you are able to please find and add it to the dataset)	
Party:	IND – Independent	TLP – Tehreek Labaik
	PMAP – Pashtunkhwa Milli Awami Party	ANP – Awami National Party
	PML-Q – Pakistan Muslim League Quaid-e-Azam	PML-N – Pakistan Muslim League Nawaz
	PPP – Pakistan Peoples Party	MQM – Muttahida Qaumi Movement
	JI – Jamat-e Islami	JUI-S – Jamiat-e Ulema Islam (Sami-ul Haq)
	JUI-P – Jamiat Ulema Islam Pakistan (F)	JUI-N – Jamiat-e Ulema Islam (Noorani)
	JWP – Jamhoori Watan Party	BNP – Balochistan National Party
	NP – National Party	GDA – Grand Democratic Alliance
	IPP – Istehkam Pakistan Party	PTI-P – Pakistan Tehreek-e Insaf Parliamentary
	PTI – Pakistan Tehreek Insaf (mixed in IND)	PAT – Pakistan Awami Tehreek
	NDM – National Democratic Movement	QWP – Qaumi Watan Party
	BAP – Balochistan Awami Party	others – All Others

Constituency	NAXXX	National Assembly Seats
	PPXXX	Punjab Assembly Seats
	PKXXX	Khyber Pakhtunkhwa Assembly Seats
	PSXXX	Sindh Assembly Seats

Votes The Votes Cast against Assembly Seat and Party

Write a python (or any language) code that is able to read the csv file and able to tokenize it cell by cell and line by line (will be used for designing map reduce tasks).

Task 1: Total Vote Casts

Write a map reduce job that is able to count the total vote casts in all seats National + Provincial Assemblies.

When done, attempt the following:

	Question	Answer
1	What was the <key,value> pair used in this query?	
2	How many mapper threads were used?	
3	How many reducer threads were used?	

4	What was the time spent by all mapper threads?	
5	What was the time spent by all reducer threads?	
6	What is the file name in which your output is located?	

Task 2: Total Vote Casts (Variation 1)

For this task, you need to calculate execution time (mapper + reducer) by two variations:

1) Play with block size of election2024.csv using the **"-D dfs.blocksize=<>"** argument.

2) Play with thread variation using the **"-D mapred.reduce.tasks=<>"**, or the **"-jobconf mapred.reduce.tasks=<>"** argument.

# of Reducer Tasks	election2024.csv block size variation (Kb)				
	4 Kb	8 Kb	16 Kb	32 Kb	Default
2					
4					
8					
16					

Highlight the best time with green and worst time with orange background.

	Question	Answer
1	How many output files are produced for 16 reducer threads.	
2	Are there some output files having 0 byte size? If so, why?	

Task 3: Total Vote Casts (Variation 2)

For this task, you need to calculate execution time (mapper + reducer) by two variations:

1) play with block size of election2024.csv using the **"-D dfs.blocksize=<>"** argument.

2) Play with thread variation using the **"-D mapred.map.tasks=<>"**, or the **"-jobconf mapred.map.tasks=<>"** argument.

# of Map Tasks	election2024.csv block size variation (Mb)				
	4 Kb	8 Kb	16 Kb	32 Kb	Default
2					
4					
8					
16					

Highlight the best time with green and worst time with orange background.

Task 4: Total Vote Casts (Variation 3)

From the Variation 1 or Variation 2, choose the election2024.csv block size which is giving best performance. And then, for this task, you need to calculate execution time (mapper + reducer) by two variations:

1) Play with thread variation using the **"-D mapred.reduce.tasks=<>"**, or the **"-jobconf mapred.reduce.tasks=<>"** argument.

2) Play with thread variation using the **“-D mapred.map.tasks=<>”**, or the **“-jobconf mapred.map.tasks=<>”** argument.

# of Map Tasks	# of Reduce Tasks			
	2	**4**	**8**	**16**
2				
4				
8				
16				

Highlight the best time with green and worst time with orange background.

Task 5: Designing Additional Queries

In this task, you have to design additional mapper/reducer threads for the following cases. In each case, it will be a good idea to put the corresponding mapper/reducer code in folders /usr/omar/task5a, /usr/omar/task5b, /usr/omar/task5c, and so on to avoid any confusion.

Task 5a: Present the Total Votes cast in each constituency.

Sample output should appear as:

NA1 – [value here]

NA2 – [Value here]

etc. etc.

Task 5b: Present the Total Votes cast in National Assembly and provinces.

Sample output should appear as:

National – [value here]

Punjab – [Value here]

etc.

Task 5c: Present the Total Votes cast against each political party in all constituencies.

Sample output should appear as:

IND – [Total Vote Value here]

PPP – [Total Vote Value here]

etc. etc.

Task 5d: The total number of candidates presented in all constituencies

Sample output should appear as:

IND – [Total Count Value here]

PPP – [Total Count Value here]

etc. etc.

Task 5e: Present the Total Votes cast against each political party in national/provincial assemblies

Sample output should appear as:

National – IND – [Value here]

National – PPP – [Value here]

etc.

Punjab – PPP – [Value here]

etc. etc.

Task 5f: Present a histogram of total vote casts in all constituencies

You already have the total votes cast in all constituencies. You now need to measure how much a certain number of votes have been cast. Sample output should appear as:

Votes between 0-25,000 – [Value here]

Votes between 25-50,000 – [Value here]

Votes between 50-75,000 – [Value here]

etc. etc.

You can experiment with the size of of histogram bins. Above sample shows 25,000 bin size.

Task 5g: Present a histogram of total vote casts against a political party in all consistuency.

You already have the total number of candidates weilded by any political party / independent in the elections. For the top three of these (in quantity), determine how much a certain number of votes have been cast in their favor. Sample output should appear as:

[name of party 1] – Votes between 0-10,000 – [Value here]

[name of party 1] – Votes between 10-20,000 – [Value here]

[name of party 1] – Votes between 20-30,000 – [Value here]

etc.

[name of party 2] – Votes between 0-10,000 – [Value here]

[name of party 2] – Votes between 10-20,000 – [Value here]

[name of party 2] – Votes between 20-30,000 – [Value here]

etc. etc.

You can experiment with the size of of histogram bins. Above sample shows 10,000 bin size.

Deliverable

Submit a report containing answers asked, along with results. It is not necessary to provide your codes. But you MUST give a description of key-value pair design used to acquire your answers.