WQD7007 Big Data Management OCC 1, 2 & 4

Lab Test (Hive)

Instructions: Work individually. Answer all the questions. Explain adequately how you get the answers that can include: 1) codes or process used; and 2) print screens and/or related files that can justify your outcome. Please create a word file in .docx to explain your answer. Submit the word file to the spectrum "Lab test" submission page.

Part 1: (8 marks)

1. Download one set of data (in .csv) about parliamentary constituencies population. Please refer to Appendix 1 (at the end of the document) on which dataset you should download. Import the downloaded dataset to HDFS. Clean the data whenever necessary. (2 marks)

2. By using Apache Hive, identify:
    a. The state that has the highest population.
    b. The year with the highest population.
    c. 5 parliament constituencies that have the lowest population among all constituencies provided.
    d. The parliament constituencies that have the highest contrast between genders.

    (6 marks)

Part 2: (7 marks)

1. Import text from the specified **web link** in Appendix 1 to HDFS (click the link). Please make sure you follow the instructions carefully.

    (1 mark)

2. Run a word count program using Hadoop MapReduce concept to count the word occurrence of the imported texts as in step 1. Save the results in HDFS. (2 marks)

3. Import the result from step 2 to Apache Hive. Display:
    a. 5 words with 5 counts in ascending alphabetical order.
    b. 10 words with lowest counts in descending alphabetical order.

    (2 marks)

4. Clean the text imported in Question 1. Then, repeat the steps in Question 2 and 3. Compare both sets of results.

    (2 marks)

Appendix 1: Please download respective datasets (in .csv format, e.g. Set1.csv) based on the name list below:

No	Matric	Name	Dataset	Weblink
1	17081572/2	M.VINOTH A/L MAHADEVAN	1	Link
2	17147301/2	MAWADDAH BINTI MUSTHAFA	2	Link
3	17167138/2	SALVIN A/L RAVINDRAN	3	Link
4	17172928/3	NURUL HAFIZAH BINTI ZAINI	4	Link
5	17202834/2	NURUL IZZAH BINTI ROSLI	5	Link
6	17203905/2	NUR MAISARAH BINTI JALALULAIL	6	Link
7	17205900/3	AHMAD MARWAN BIN MURSHIDI	7	Link
8	17206574/2	TAN RIK EE	8	Link
9	17206926/2	MUHAMAD RAFIQ IQBAL BIN SAMSUDIN	9	Link
10	17207012/2	TAN JIAN LIN	10	Link
11	22091346/1	TEO KAI NING	1	Link
12	22102113/1	YANWEI ZHANG	2	Link
13	22106713/1	LI YUE XIN	3	Link
14	23063431/1	CHANG QI HAN	4	Link
15	23074766/1	TAN FOO HOU	5	Link
16	23075339/1	WU YIBIN	6	Link
17	23079722/1	MUHAMMAD HAKIM BIN NASARUDDIN	7	Link
18	23083543/1	FOO KAI WEN	8	Link
19	23083896/1	CHEAH YINING	9	Link
20	23086179/1	LIM WEI EN	10	Link
21	23087440/1	LEE MIN QI	1	Link
22	23092124/1	LOW E-JIE	2	Link
23	23092758/1	LOH KE YI	3	Link
24	23094470/1	CHUA SZE YAN	4	Link
25	23095937/1	LAU WEN XI	5	Link
26	23096526/1	NUR ARIANA SOFEA BINTI BADRUL HISHAM	6	Link
27	23097292/1	SHAIK MOHAMMED MUZEEB	7	Link
28	23101507/1	GE,ZHIXING	8	Link
29	23101520/1	CHAN YUNG HER	9	Link
30	23101857/1	VETRI A/L THANABALAN	10	Link
31	23102307/1	LUO QINGZHEN	1	Link
32	23102772/1	LIM SZE GEE	2	Link
33	23104045/1	LISA HO YEN XIN	3	Link
34	23104111/1	CHOW KOO LI	4	Link
35	23104958/1	RAHAYU RIANTI	5	Link
36	23105612/1	TOH CHU XIAN	6	Link
37	23107324/1	HUANG LILI	7	Link
38	23108481/1	WONG MEI KAIT	8	Link
39	23108677/1	ZHAO ZITONG	9	Link
40	23109154/1	ESTHER TEO YONG XUAN	10	Link
41	23110707/1	NICHOLAS OOI JIAWEI	1	Link
42	23111363/1	WANG SHIBO	2	Link
43	23117259/1	BAIZID YALDRAM	3	Link
44	23117951/1	KARENINA KAMILA	4	Link
45	23121085/1	KEE SHAO CHONG	5	Link
46	23121328/1	MOHAMMED IQRAM	6	Link
47	23121442/1	CLEMENT LIM KEE MIN	7	Link
48	23122622/1	XIAN ZHIYI	8	Link
49	24050548/1	LEE XUAN YU	9	Link
50	24053855/1	LOOI XUE YING	10	Link
51	24055123/1	PEE JIUN BIN	1	Link
52	24056266/1	YANG JIYU	2	Link

53	24056475/1	LIU YILIN	3	Link
54	24056751/1	AMIRAH ILHAMI BINTI AZIZUL	4	Link
55	24057531/1	ANG ZHI YANG	5	Link
56	24057547/1	ZHANG XIAOMENG	6	Link
57	24057850/1	SHI YAN	7	Link
58	24059242/1	LOONG SHIH-WAI	8	Link
59	24060333/1	WANG JIAJU	9	Link
60	24060558/1	YEO MINYI	10	Link
61	24061230/1	NGO XING YING	1	Link
62	24064261/1	AHMAD DANIEL BIN MOHD SUPANDI	2	Link
63	S2110638/1	KARAM ALJANADI	3	Link
64	S2119226/2	MUHAMMAD TAUFIQ BIN ISMAIL	4	Link
65	S2152880/1	WONG YI TING	5	Link
66	S2157250/1	LIM SZE CHIE	6	Link
67	U2004540/2	DHIVIYANANDHINI A/P V KUMAR	7	Link
68	U2004720/2	MOHAMMAD IQBAL AFIF BIN MOHAMAD SHAHNAZ	8	Link
69	U2004763/2	KHOR KEAN TENG	9	Link
70	U2005486/1	AFRINA ROSA BINTI MOHAMAD SATTAR	10	Link
71	17127309/2	MD AZRUL SYAFFIQ BIN MD SUHAIMI	1	Link
72	17137568/2	MOHD NOOR SYAWAL FAQQIE BIN ABDULLAH	2	Link
73	17196177/2	LOH EE SAM	3	Link
74	17204762/2	HAU JIA QI	4	Link
75	17207347/2	AHMAD KAMIL HARIZ BIN HARIZAL	5	Link
76	23052209/1	QIU YIMEI	6	Link
77	23082194/1	ARUN KUMAR	7	Link
78	23086480/1	ANG ENG HOOI	8	Link
79	23090576/1	TAN SIAO SHUEN	9	Link
80	23090797/1	PANG SUK MIN	10	Link
81	23091969/1	TEOH SUE LYNN	1	Link
82	23094825/1	SITI HAIRUNEESHA BINTI ZAINUDDIN	2	Link
83	23098984/1	LIU HUI	3	Link
84	23100971/1	ONG HUI LING	4	Link
85	23101952/1	ZHANG YUMAN	5	Link
86	23103507/1	ZOU JINGYI	6	Link
87	23104578/1	XIAO WANJUN	7	Link
88	23105890/1	TAN CHEE YONG	8	Link
89	23106248/1	LAW YU XUAN	9	Link
90	23108894/1	LIJINZHAO	10	Link
91	23113750/1	CHENHUILIN	1	Link
92	23117091/1	WU XIAOYUE	2	Link
93	23118928/1	GOH WEE KEN	3	Link
94	23119264/1	AREEJ ABDURAHMAN MOHAMMAD	4	Link
95	23122060/1	DONG QIYUAN	5	Link
96	24052516/1	LI JUNMING	6	Link
97	24054402/1	SITI NURAISHAH BINTI AB MANAM	7	Link
98	24056080/1	TIAN TIANCHU	8	Link
99	24056483/1	YAO YUANWEI	9	Link
100	24059144/1	MUHAMAD AIMAN BIN JAMAL ABD NASIR	10	Link
101	24059665/1	YANGYANG LYU	1	Link
102	24064883/1	WANG KEXIN	2	Link
103	S2028822/1	PAARISHA EMILIE	3	Link
104	17072752/2	MELANIE NG HUEI YEE	4	Link
105	17083819/2	KEVIN WONG XIN KAI	5	Link
106	17127681/2	TAN YONG SHENG	6	Link
107	22075938/1	NITYA A/P PONNUSAMY	7	Link

108	22080395/1	LEE QIAN HUI	8	Link
109	22082745/1	HASSAN HASSANZADEH ALIABADI	9	Link
110	22090924/1	YUKI SIM TZE YII	10	Link
111	22093309/1	LEGAWATTHI A/P THIYAGARAJAN	1	Link
112	22098312/1	JIANG JIAYAN	2	Link
113	22104060/2	ZHANG WEI	3	Link
114	22107573/1	IZZUL ILHAM BIN YUSOF	4	Link
115	22109241/1	LAI YONG LIN	5	Link
116	22111473/1	RABITA BHUIYA TANAYA	6	Link
117	22120484/1	HUA SHAOJIE	7	Link
118	23060772/1	AIDA WANIE BINTI JASNI	8	Link
119	23062959/1	ALI BIN ABDUL RAHMAN	9	Link
120	23063305/1	LOW MENG FEI	10	Link
121	23068710/1	YAP J-SHYNE	1	Link
122	23069166/1	CHOO WAN QI	2	Link
123	23070711/1	CHONG KAH HOE	3	Link
124	23072236/1	SIM JIN XIANG	4	Link
125	23076263/1	SHARIFAH NURUL AMIRAH BINTI SYED AHMAD FAUZI	5	Link
126	23085604/1	JOANNE CHANG YII	6	Link
127	23086487/1	LOW ZI YANG	7	Link
128	S2176972/1	MUHAMMAD NASRULLAH BIN MOHAMED BUNYAMIN	8	Link
129	S2193226/1	KAI ERN CHOW	9	Link