VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING

Mid-Term Examination Summer Semester MCA - 2022-2023

Course Name: BigData Analytics **Duration: 1hr 30Mins.**

Course Code: ITA6008 **Max. Marks : 50 marks**

Slot : C1+TC1+C2+TC2 Faculty : Prof. Pounambal M, Prof. Balasubramani and Prof. Rajiv

---

Q.No

Answer ALL Questions 5 * 10 = 50 Marks

1 Specify the Industry vertical under which the following are categorized. Write one challenge and application for each.

a. Spotify [0.5+1+1]

b. Carnegie Learning [0.5+1+1]

c. City of Dublin [0.5+1+1]

d. Smart Grid [0.5+1+1]

2 Derive the data analytics life cycle for recommending an item in AMAZON website.

3 a. A file named moviedb.txt of size 500KB is stored in HDFS. A client wants to read the file. Explain the components involved in the above scenario. [5]

b. Write the features of the following type of data, classify the type of data also guide the user to store it in HADOOP environment. [2.5+2.5]

i)

```
<CATALOG>
<PLANT>
<COMMON>Bloodroot</COMMON>
<BOTANICAL>Sanguinaria canadensis</BOTANICAL>
<ZONE>4</ZONE>
<LIGHT>Mostly Shady</LIGHT>
<PRICE>$2.44</PRICE>
<AVAILABILITY>031599</AVAILABILITY>
</PLANT>
<PLANT>

<COMMON>Jack-In-The-Pulpit</COMMON>
<BOTANICAL>Arisaema triphyllum</BOTANICAL>
<ZONE>4</ZONE>
<LIGHT>Mostly Shady</LIGHT>
<PRICE>$3.23</PRICE>
<AVAILABILITY>020199</AVAILABILITY>
</PLANT>
```

<PLANT>

<COMMON>Cardinal Flower</COMMON>
<BOTANICAL>Lobelia cardinalis</BOTANICAL>
<ZONE>2</ZONE>
<LIGHT>Shade</LIGHT>
<PRICE>$3.02</PRICE>
<AVAILABILITY>022299</AVAILABILITY>
</PLANT>
</CATALOG>

ii)

Persons			
PersonID	Name	Age	Salary
P1	John	32	$400000
P2	Johnny	33	$410000
P3	Janet	31	$400000
P4	Jeremy	32	$450000
P5	Justin	33	$600000
P6	Jazmyn	35	$250000
P7	Judy	30	$900000
P8	Jolly	33	$100000
P9	Jack	31	$120000

4 If you have rights to access VTOP access login permission also you need to collect the 2nd and 3rd year student information data of all VIT branches from VTOP and need to be stored in Hadoop echo system, to do so identify the suitable tools available in Hadoop ecosystem and brief on it.

5 Explain the Map reduce process supports to count the total number of occurrences of each single word present in text document

VIT®
Vellore Institute of Technology

Summer Term II Final Assessment Test – August 2023

Course: ITA6008 - Big Data Analytics

Class NBR(s): 1047/1049/1086 Slot: C1+TC1+C2+TC2

Time: Three Hours Max. Marks: 100

KEEPING MOBILE PHONE/SMART WATCH, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE

Answer any TEN Questions
(10 X 10 = 100 Marks)

1. a) Compare Business Intelligence and Data science with required examples. [5]

   b) Describe the challenges involved in the current analytical architecture from the perspective of data scientists. [5]

2. Describe the architecture of HDFS with neat diagram and explain any four components.

3. Write applications to process the data regarding the electrical consumption of an organization which contains the monthly electrical consumption and the annual average for various years and produce results such as finding the year of maximum usage, year of minimum usage. Describe the steps to execute mapreduce program using java.

4. Illustrate how data is stored in a Hadoop environment with master nodes and *worker nodes*?

5. a) Explain How google file system differs from the hadoop file system? [5]

   b) Explain with a neat sketch the processing of a job in Hadoop/map reduce. [5]

6. An enterprise has 500 GB of unstructured data to be processed. Suggest a suitable architecture for the data processing work. Explain with a diagram, all the components of the file system supported by your solution.

7. List the various operational modes of hadoop cluster configuration and explain in detail about configuring/installing the hadoop in local/standalone mode. What is the use of each mode from application and developer point of view?

8. Demonstrate exploratory data analysis using R with suitable example data set.

9. Apply Wilcoxon Rank-Sum Test procedure to the below scenario.

A production planner want to see if the operating rates for 2 factories is the same. For factory 1, the rates are 71, 82, 77, 92 & 88. And for factory 2, the rates are 85, 82, 94 & 97. Assume the level of significance as 5%. Refer Fig.1. For critical values.

$\alpha = 0.025$ one-tailed; $\alpha = 0.05$ two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

$\alpha = 0.05$ one-tailed; $\alpha = 0.10$ two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

Fig.1. Wilcoxon Rank Sum Table

10. Describe Hive components with a neat diagram.

11. What is the significance of Apache pig in Hadoop context? Describe the main components and the working of Apache Pig with a simple example.

⇔⇔⇔