Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First Semester 2023-2024

Comprehensive
(EC3M Examination)
(Sample Question Paper)

Course No. : DSECLZG522
Course Title : Big Data Systems
Nature of Exam : Open Book – Typed Only
Weightage : 40%
Duration : 2.5 Hours
Date of Exam :

No. of Pages = 3
No. of Questions = 8

Note to Students:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

---

1. What limitation of Hadoop is getting addressed in HBase? HBase persists data in HTables in HDFS and Cassandra stores data in SSTables. HTables and SSTables are immutable. Explain how updates are deletion of rows is implemented in HBase and Cassandra?

(5 Marks)

2. A company decided to perform real time analytics on its Web server logs to gain insights on website visitors, behavior, crawlers accessing the site, business insights, security issues, and more. For this, the log file entries are to be streamed to an HDFS folder in a Hadoop cluster. Streaming the logs to HDFS is to be done only when 1000 or more entries are made to the webserver log. Configure a Flume agent to implement the streaming with webserver log as the source and an HDFS folder as the sink.

(5 Marks)

3. Carefully go through scala Code1 and scala Code2 give below and answer the questions given below them.

Spark Scala Code1

```
val lines = new Array[String](2)
lines(0) = "Hello world"
lines(1) = "How are you world"
val stringRDD=sc.parallelize(lines,1)
val wordsRDD = stringRDD.flatMap(x => x.split(" "))
val wordRDD = wordsRDD.collect()
wordRDD
```

Spark Scala Code2

```
val lines = new Array[String](2)
lines(0) = "Hello world"
lines(1) = "How are you world"
val stringRDD=sc.parallelize(lines,1)
val wordsRDD = stringRDD.flatMap(x => x.split(" "))
val filtRDD = wordsRDD.filter(x => x.startsWith("H"))
val wordRDD = wordsRDD.collect()
wordRDD
```

Question 1. - Will there be any difference in the execution time for Code1 and Code2 ?
Question 2. - Modify Code1 to run as 2 tasks in parallel
Question 3. - Modify the Code2 to output only words having a length of 5
Question 4. - Modify Code2 to output the number of unique words in the list.

(5 Marks)

4. Describe the differences between the following 2 tables defined in Cassandra NoSQL database from the point of view of storage and retrieval of the data from the tables. In which scenario, retrieving data from Table.1 will be faster than retrieving data from Table.2

Table.1

```
CREATE TABLE application_logs (
   id              INT,
   app_name  VARCHAR,
   hostname   VARCHAR,
   DateTime   TIMESTAMP,
   env            VARCHAR,
   Log_level   VARCHAR,
   log_msg     TEXT,
   PRIMARY KEY ((app_name, env), hostname)
);
```

Table.2

```
CREATE TABLE application_logs (
   id              INT,
   app_name  VARCHAR,
   hostname   VARCHAR,
   DateTime   TIMESTAMP,
   env            VARCHAR,
   Log_level   VARCHAR,
   log_msg    TEXT,
   PRIMARY KEY ((app_name), env, hostname)
);
```

(5 Marks)

5. Why windowing is needed in Stream computing? What are the 3 different types of windowing methods used in Spark Streaming Analytics? Explain the difference between these methods.

An IOT device transmits 1 event in every 10 seconds. The Stream Analytics system need 10 consecutive events to make an inference. The system should output one inference in every minute. Develop a suitable windowing scheme to meet these requirements.

(5 Marks)

6. You are designing a Distributed Hash Table to store 128 data values in a table distributed on a 5- node cluster. The primary key consists of 8 characters consisting of 7-bit ASCII codes.
(1) Design a hashing algorithm for equally distributing data in the tables on the nodes of the cluster.
(2) What percentage of data will get stored on one node.
(3) Given below are 2 schemes for distributing the tokens generated from the partition keys on the 5 nodes of the cluster:
Scheme 1:
Node0– 0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125
Node1 - 1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126
Node2 -2 7 12 17 22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107 112 117 122 127
Node3 -3 8 13 18 23 28 33 38 43 48 53 58 63 68 73 78 83 88 93 98 103 108 113 118 123

Node4 - 4 9 14 19 24 29 34 39 44 49 54 59 64 69 74 79 84 89 94 99 104 109 114 119 124

Scheme 2 :
Node 0 – 0 to 25 (both inclusive)
Node 1 – 26 to 51 (both inclusive)
Node 2 – 52 to 77 (both inclusive)
Node 3 – 78 to 103 (both inclusive0
Node 4 - 104 to 127 (both inclusive)
Which of the scheme mentioned above will be giving almost equal distribution of tokens on all nodes of the cluster? Give justification to your answer.

(5 Marks)

7. A spark cluster is to be configured to process a file of size 48 GB. The processing is of CPU intensive nature. You are asked to configure the cluster on a 5-node cluster in which 1 node will have to be used as the driver and the other 4 nodes will have to be used as worker nodes. Each of the nodes of the cluster is having 16 Gib memory and 16 core CPUs. On each of the nodes, you need to reserve 4Gib for the operating system and other applications. The rest of the memory can used for Spark. On each of the worker nodes, you need reserve 4 cores for the operating system and other processes. For optimum performance, give values for the following:

1. Number of RDD partitions
2. Number of executors
3. Number of cores and size of memory allocated to each executor
4. Number of tasks getting executed on each executor

(5 Marks)

8. You have a 16384 MB file stored on HDFS as part of a Hadoop 3.x distribution. A data analytics program stores this file on the HDFS cluster with 3 data nodes and runs in parallel across the cluster nodes.

(a) The default values for HDFS block size and the replication factor is used in the configuration. How many total blocks of the data file including replicas will be stored in one node of the cluster?

(b) The cluster has 48 cores to speed up the processing. If the program can at best achieve 60% parallelism in the code to exploit the multiple cores and the rest of it is sequential, what is the theoretical limit on speed-up you can expect with 48 cores compared to a sequential version of the same program running on one core with the same file? How will this limit change if you doubled the compute power to 96 cores? You can simplify the system to assume cluster nodes and cores mean the same and we can ignore the overheads of communication etc. depending on the specific cluster configuration, scheduling etc.

(c) Suppose you could use a more scalable algorithm with 80% parallelism and a larger file as you move to a 108-core system. What would be the theoretical speed-up limit for 108 cores?

(5 Marks)

****