**Birla Institute of Technology & Science, Pilani**
**Work Integrated Learning Programmes Division**
**First Semester 2024-2025**

**Mid-Semester Test**
**(EC-2 Regular)**

| | |
|---|---|
| Course No. | : DSECLZG522 |
| Course Title | : Big Data Systems |
| Nature of Exam | : Closed Book |
| Weightage | : 30% |
| Duration | : 2 Hours |
| Date of Exam | : 18-01-2025 (EN) |

No. of Pages = 2
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.

Q.1. In a distributed computing system consisting of multiple nodes, data is accessed by an application from other nodes. The data is cached in a memory of size 10MB on the local node, and it can be delivered only from this cache to the application. Calculate the average access time of a block of data consisting of 1MB fetched from one node and processed sequentially from memory in another node of the cluster if the hit rate is 80%. You may ignore the advantages in the access time contributed by the filesystem / network caching. Typical timing values are given below:

To read 1MB Sequentially from disk = 20,000 μs
To read 1MB sequentially from memory = 250 μs
To send 1MB over 1Gbps network = 100 μs

You need to create a suitable relationship between the parameters and calculate the average access time using the formula created by you.

[ Tavg = h * Tm + ( 1-h ) * ( Td + Tn + Tm ) ]

[3 Marks]

Q.2. In a data engineering pipeline, a computation task processes data using parallel execution across multiple nodes. The system's speedup is modelled using the **Gustafson-Barsis Law**, which considers a problem's scalability by increasing the workload proportionally to the number of processors. Given the following parameters:

- Fraction of the workload that is inherently sequential: 0.15
- Number of processors: 32

Calculate:

(a) The achievable speedup of the system as per Gustafson-Barsis Law.
(b) If the number of processors is increased to 64 while maintaining the same sequential workload fraction, how much improvement in speedup is expected?

[3 Marks]

Q.3. A web application consists of 2 application servers sharing the load. The data is persisted with 1 database server. The application servers have an MTBF of 1 year and an MTTR of 12 hours. For the database server, the MTBF is 3 years with an MTTR of 1 week. Calculate the combined availability of the total system.

[3 Marks]

Q.4. Explain MongoDB read concerns and its types?

[3 Marks]

Q.5. List the types of parallelism and explain the type of parallelism used in MapReduce.

[3 Marks]

Q.6. What is meant by Consistency in CAP theorem? Let’s establish a few definitions:
$N$ = The number of nodes that store replicas of the data.
$W$ = Number of replicas that need to acknowledge the receipt of update before update completes.
$R$ = Number of replicas that are contacted when a data object is accessed through a read operation.
What type of consistency can be guaranteed in the following scenarios?

(a) $W+R > N$
(b) $N=2$, $W=2$, and $R=1$
(c) $N=2$, $W=1$, and $R=1$

[3 Marks]

Q.7. An automobile manufacturing company is implementing a solution to automate their manufacturing process. Each machine used in the manufacturing process contains one or more sensors that captures data such as temperature, speed of a moving part, etc. They have decided to use Cassandra database for storing high volume data originating from sensors in the machines. This data is going to be stored in a Cassandra cluster that you must design and maintain. The Cassandra cluster is setup in the same rack in the same data center. Once the application is up and running, the data it captures will be provided to analytics tool chains. The application is write-intensive; tens of millions of events will be generated per second, and the company has an SLA requiring that all events be persisted across multiple nodes in the Cassandra cluster in less than 10ms.

(a) Suggest values for class and replication factor for the Key Space
(b) Suggest suitable Write consistency level for persisting data into the cluster
(c) Suggest a suitable PRIMARY KEY (Primary and clustering columns) for the table given below:

```
CREATE TABLE sensor_data (
          serial_number text,
          date text,
          snapshot_time timestamp,
          facility_id int,
          sensor_type text,
          sensor_value text,
);
```

[3 Marks]

Q.8. You have a 928 MB file stored on HDFS as part of a Hadoop 2.x distribution. A data analytics program uses this file and runs in parallel across the cluster nodes. Default block size and replication factor are used in the configuration. (a) What will be the total number of blocks including the replicas that will be stored in the cluster? (b) What are the unique HDFS block sizes you will find for the specific file? (c) Why large block sizes are selected for HDFS?

[3 Marks]

***********