Big Data Analytics
Assignment-02 (Spring 2024)

Submission to be done on portal

Due Date: Friday, May 3, 2024, before 11:59 PM

- **Use Assignment pages and scan your assignment. (Hand written Only)**
- **Upload .pdf file on portal**
- **[Note: Title page must have Student Name, Registration Number, Section, and Date of Submission.**
- **Don't use register pages**
- **`Submission will be on portal only`**

<table>
<tr><th colspan="3">Assignment 02</th></tr>
<tr><td>Q #01:</td><td>How does the implementation of erasure coding (EC) in HDFS, address the inherent storage overhead of replication while maintaining durability guarantees, and what are the key design considerations and optimizations made to seamlessly integrate EC into the existing distributed storage system of HDFS, encompassing modifications to the NameNode, DataNode, and client read and write paths,what future developments are anticipated regarding support for varied data layouts and advanced EC algorithms?</td><td>5 Marks</td></tr>
<tr><td>Q #02:</td><td>How do replication schemes like RAID-1 and erasure coding (EC), particularly Reed-Solomon (RS), differ in their approaches to tolerating disk failures, and how do their configurations impact data durability, storage efficiency, and storage overhead, considering factors such as the number of tolerated simultaneous failures and the ratio of data cells to parity cells?
(Hint: You can use following table to write answer of question #02)
<table>
<tr><th></th><th>Data Durability</th><th>Storage Efficiency</th></tr>
<tr><td>Single replica</td><td>0</td><td>100%</td></tr>
<tr><td>Three-way replication</td><td>2</td><td>33%</td></tr>
<tr><td>XOR with six data cells</td><td>1</td><td>86%</td></tr>
<tr><td>RS(6,3)</td><td>3</td><td>67%</td></tr>
<tr><td>RS(10,4)</td><td>4</td><td>71%</td></tr>
</table>
</td><td>5 Marks</td></tr>
</table>