COURSE CODE (CREDITS): 22M11CI213 (3)

COURSE NAME: Big Data Analytics

MAX. MARKS: 25

COURSE INSTRUCTOR: Deepak Gupta

MAX. TIME: 1:30 Hr.

Note: 1) All questions are compulsory. Marks and COs are indicated against each question. 2) Attempt questions in the given sequence. 3) Be precise in your answers. 4) Write neatly.

Q 1. Consider a set of 5 pages with initial rank 1/5 and apply Page Rank algorithm to calculate the final page rank of each page. **[5] CO2**

Q 2. Draw a block diagram to represent driver program, cluster manager, worker nodes, executor and tasks in Apache Spark. Also briefly describe their usage by taking a suitable example. **[5] CO2**

Q 3. (a) Differential between data lake and data warehouse with suitable examples.
(b) Comparison between Apache Spark, Apache Flink and Apache Storm. **[2.5 *2] CO2**

Q 4. Consider a log mining use case in which error messages are loaded into memory and thereafter, searched for finding useful insights/patterns. Implement Apache Spark RDDs, transformations and actions in Scala/Python/Java to accomplish the given use case. **[5] CO3**

Q 5. (a) Describe any five functions used to accomplish Apache Spark transformations and actions.
(b) Write a program using Spark SQL to find number of teenagers in a given list of persons. **[2.5 *2] CO3**