**MANAV RACHNA INTERNATIONALINSTITUTE OF RESEARCH AND STUDIES (DEEMED TO BE UNIVERSITY)**

Faculty of Engineering and Technology (FET)

**Department of Computer Science and Engineering**

**Tutorial/Assignment Sheet**

**Subject Name: BDA Tutorial No.: 2**

**Subject Code: Class / Semester: VII**

**Max. Marks : 50M Time Allowed: 1 Week**

|  |  |  |
| --- | --- | --- |
| **1.**  **2.**  **3.**    4**.**    **5.**      **6.**    **7.**  8.  9.  10. | Write an overview of Hadoop by citing its brief background and five key aspects of Hadoop.  What are the four modules that make up the Apache Hadoop Framework? Draw the framework in Block Diagarm for Hadoop 1.0 and Hadoop 2.0.  Define a cluster and further explain NameNodes and DataNodes in a Hadoop Cluster. Mention the number of NameNodes and DataNodes that can run on a single Hadoop Cluster.  Calculate the data storage in Hadoop cluster for the given input:  Suppose Hadoop cluster is getting 120 TB data on daily basis and we have default replication factor as 3, calculate the daily data storage requirement.  **Hint:** Storage is based on formula {Storage = Daily data ingestion\*Replication}  Calculate Number of Data Nodes for the **above scenario:**  Suppose we have servers with " just a bunch of disks"(JBOD) of 10 disks and each disk is having 4 TB storage size.  **Hint:** No of Data Nodes =( Daily data ingestion/data node capacity )  Differentiate between the following:  a) Big Table and HDFS  b) Replication and Sharding  **c**) PetaByte vs. BrontoByte Three approaches to improving the performance by orders of magnitude are scale-down, Scale-up and Scale-out for Hadoop environment. Is scale-out a right approach for Hadoop’s performance, why? Define a Daemon and mention the various Hadoop Daemons with their roles in a Hadoop cluster.   1. What was Hadoop written in? a) Java (software platform) c) Java (prog. language) b) Perl d) Lua (prog. language) 2. Which of the following platforms does Hadoop run on? a) Bare metal c) Cross-platform b) Debian d) Unix-like 3. Hadoop achieves reliability by replicating the data across multiple hosts, and hence does not require \_\_\_\_\_\_\_\_ storage on hosts. a) RAID c) ZFS b) Standard RAID levels d) Operating system 4. The Apache MAHAUT project’s goal is to build a scalable………………..tool. 5. Hive is not a relational database, but a …………..that supports the parts of SQL specific to querying large datasets stored in Hadoop-compatible file systems. 6. \_\_\_\_\_\_\_\_\_\_\_ is general-purpose computing model and runtime system for distributed data analytics. 7. The ………….. Scripting language is not only a higher-level data flow language but also has operators similar to SQL. | **COs:**  CO2  CO2  CO2  CO2  CO2  CO2  CO2  CO2  CO2  CO2 |

**Prepared By: Approved By:**

**Dr. Tapas Kumar**

**Course Coordinator (Professor &HOD-CSE)**