Exam Slot → **A**

APJ Abdul Kalam Technological University
Third Semester M.Tech Degree Examination December 2016
Ernakulam II Cluster
COMPUTER SCIENCE AND ENGINEERING
Specialization: COMPUTER SCIENCE AND ENGINEERING

05CS 7041- BIG DATA PROCESSING (ELECTIVE- IV)

Time: 3 hrs. Max. Marks: 60

I a) The problem with the current HFile format is that it causes high memory usage and slow up times for the region server because of large bloom filters and block index sizes. Write a note on the new feature that has been added to Hbase to overcome this problem. (4 Marks)

b) In Hadoop, how can we store a large amount of data, and provide access to this data to many clients distributed across a network. Illustrate the architecture of Hadoops implementation of the distributed file system. (8 Marks)

II a) Explain with an example how MapReduce can be implemented for an image processing system. (6 Marks)

b) Why Fibonacci series cannot be implemented using MapReduce. (6 Marks)

III a) Write the Hive queries for the following scenario: Create a table employee as external table which partitioned based on date of joining(doj) with the following fields :
(id INT, name STRING, department STRING, salary INT, doj STRING). (7 Marks)

Each field will be separated by a delimiter ‘,’. Add two partitions for 2-02-2010 and 25-10-2011 into the hdfs under master- server/user/employee. Copy this to the location in ‘amazon S3n://outbucket/users‘and point it to s3 location. Remove the Hfds copy after this. Display the partitions and the external table information along with the schema information of the employee table.

	b)	Hive provides the high level abstraction for Hadoop system and the programmer need not worry about the low level details. Discuss how Hive achieves this by clearly indicating the components and their interaction.	(7 Marks)
	c)	Examine the distinct features that will differentiate pig from map-reduce.	(4 Marks)

OR

IV	a)	Discuss the concept of combining and splitting the data.	(6 marks)
	b)	Explain in detail about creating, altering, partitioning and managing tables in hive.	(12marks)
V	a)	Assume that you are working in an agency whose main focus is on weather forecasting implemented on a Hadoop cluster environment. There exists a wide variety of real time data to be pushed to the cluster from various satellites. So there is a need for scalable, high-throughput, fault-tolerant processing of these data to predict the weather. Suggest how you can achieve this with the help of spark by describing the process and figures.	(9 Marks)
	b)	Automated starting of the workflow process is done in oozie server with the help of certain functional component. Identify this component and describe the key concepts and language elements that help its execution.	(9 Marks)

OR

VI	a)	How does Spark provide its speed and avoid Disk I/O while retaining the attractive fault tolerance, locality and scalability properties of MapReduce.	(9 Marks)
	b)	Explain about the spark architecture	(9 Marks)