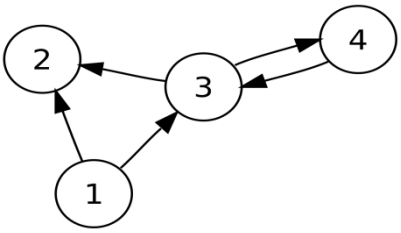
VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Model Final Assessment Test (FAT) Question Paper

Programme	:	**M.Tech**	Semester	:	**Fall 2019 – 20**
Course Title	:	**Bigdata Frameworks**	Code	:	**CSE6001**
			Slot	:	**C1**
Faculty	:	**Prof. Ramesh Ragala**	Class Nbr	:	
Time	:	**Three Hours**	Max. Marks	:	**100**

Answer any ALLquestions

Q.No.	Sub. Sec.	Question Description	Marks
1.		List down the drawbacks of MapReduce programming model. Illustrate any one drawback of MapReduce programming with suitable example.	**10**
2.		Provide the architectural framework of Hadoop, Spark and Spark Streaming. Compare and constract the features of each framework.	**12**
3.		Illustrate the working procedure of MapReduce framework with your own example. Explain map and reduce steps diagrammatically.	**10**
4.		 The above directed graph represents connections among the twitter user. The vertices represent the twitter user. An edge represents the *"following"* relation among the users. An incoming edge indicates a "follower". Write apache spark program to disply the followers of every user.	**10**
5.		A Data node in a Hadoop cluster has following configuration: Each data node has 10 disks and each disk has the storage capacity of 1TB. Each data node uses 2 disks to store operating system, logs, etc. Case – I: The user wants to store 600TB data with out any compression into a Hadoop Distributed cluster. Calculate total number of data nodes required to store specified data. Assume intermediate and replication factors as 1/4 and 3 respectively. Case – 2: The user want to store compressed data with compression ratio as 2.3 into Hadoop cluster. Whereas the disk space utitlization in data nodes after compression is 65%. Calculate the total number of required data nodes. Assume intermediate and replication factors as 1/4 and 3 respectively.	**10**
6.		Analytic Vidya Pvt Ltd has developed social search and advertising application, which gives recommendations based on search history of user and his/her friends in Google, Bing, eBay, amazon, wikipedia. This application will be generating 150ZB of logs and publish 17 millions of digital advertisements everyday. After adding a new feature to this application, the data generation has crossed 300ZB everyday. Analyse and processing of this humongous data on small Hadoop cluster give abnormal behavior like node failure, unable to identify replica nodes, etc. How do you solve this issue and justify your answer.	10

<table>
<tr><td>7.</td><td></td><td>Illustrate the different data – oriented processing tools on Hadoop frawework.</td><td>10</td></tr>
<tr><td>8.</td><td></td><td>User-information Dataset has attributes like UserID, e-mail, language and address. Transaction-Information dataset has transID, productID, userID, amount and ItemID as attributes. As a Data Scientist, find the number of unique locations in which each product has been sold using MapReduce application. Hint: join the two datasets together</td><td>10</td></tr>
<tr><td>9.</td><td></td><td>Illustrate the need of streaming analytics in big data era and explain the role of Spark Streaming with your own example.</td><td>10</td></tr>
<tr><td>10</td><td></td><td>Ravi and Rakesh have scored marks in M1,M2 and M3 courses. The following table gives the details of the marks.
<table>
<tr><th>Name of student</th><th>M1</th><th>M2</th><th>M3</th></tr>
<tr><td>Ravi</td><td>85</td><td>75</td><td>95</td></tr>
<tr><td>Rakesh</td><td>65</td><td>45</td><td>85</td></tr>
</table>
Write Apache spark application to disply list of marks of the Ravi and Rakesh based on each subject as a record like (M1,(85,65)).</td><td>8</td></tr>
<tr><td></td><td></td><td>Total marks</td><td>[100]</td></tr>
</table>

⇔⇔⇔