

Program: **BE EXTC** Engineering  
Curriculum Scheme: **Revised 2016**  
Examination: **Final Year Semester VII**

Course Code: **ECCDLO7032** and Course Name: **Big Data Analytics**

Time: 1 hour

Max. Marks: 50

=====

=====

Note to the students:- All the Questions are compulsory and carry equal marks .

Q1.	DataStage originated at _____, a company that developed two notable products: UniVerse database and the DataStage ETL tool
Option A:	Vmark
Option B:	Vzen
Option C:	Hatez
Option D:	Smark
Q2.	The feature of big data that refers to the quality of the stored data is _____
Option A:	Variety
Option B:	Volume
Option C:	Variability
Option D:	Veracity
Q3.	Data locality feature in Hadoop means
Option A:	<u>store the same data across multiple nodes.</u>
Option B:	<u>relocate the data from one node to another.</u>
Option C:	<u>co-locate the data with the computing nodes.</u>
Option D:	Distribute the data across multiple nodes.
Q4.	Job tracker runs on
Option A:	Namenode
Option B:	Datanode
Option C:	Secondary namenode
Option D:	Secondary datanode
Q5.	Pig latin is a:
Option A:	SQL language
Option B:	Database language
Option C:	Dataflow language
Option D:	Programmimg language
Q6.	The _____ executes the Mapper/ Reducer <i>task</i> as a child process in a separate jvm.

Option A:	TaskTracker
Option B:	reducer
Option C:	avro
Option D:	mapper
Q7.	The fundamental idea of _____ is to split up the functionalities of resource management and job scheduling/monitoring into separate daemons
Option A:	Yarn
Option B:	reducer
Option C:	mapper
Option D:	scheduler
Q8.	The _____ is the ultimate authority that arbitrates resources among all the applications in Yarn system
Option A:	ResourceManager
Option B:	Scheduler
Option C:	Nodemanager
Option D:	systemmanage
Q9.	<i>Which of the following is NOT supervised learning?</i>
Option A:	PCA
Option B:	Decision Tree
Option C:	Linear Regression
Option D:	Naive Bayesian
Q10.	<i>Which of the following statements about Naive Bayes is incorrect?</i>
Option A:	Attributes are equally important
Option B:	Attributes are statistically dependent of one another given the class value
Option C:	Attributes are statistically independent of one another given the class value
Option D:	Attributes can be nominal or numeric
Q11.	<i>Suppose we would like to perform clustering on spatial data such as the geometrical locations of houses. We wish to produce clusters of many different sizes and shapes. Which of the following methods is the most appropriate?</i>
Option A:	Decision Trees
Option B:	Density-based clustering
Option C:	Model-based clustering
Option D:	K-means clustering
Q12.	What is the recommended best practice for managing big data analytics programs?
Option A:	Adopting data analysis tools based on a laundry list of their capabilities
Option B:	Letting go entirely of "old ideas" related to data management

Option C:	Focusing on business goals and how to use big data analytics technologies to meet them
Option D:	The ability of business intelligence and analytics vendors to help them answer business questions in big data environments
Q13.	Companies that have large amounts of information stored in different systems should begin a big data analytics project by considering
Option A:	The creation of a plan for choosing and implementing big data infrastructure technologies
Option B:	The interrelatedness of data and the amount of development work that will be needed to link various data source
Option C:	The ability of business intelligence and analytics vendors to help them answer business questions in big data environments
Option D:	Adopting data analysis tools based on a laundry list of their capabilities
Q14.	Cloudera Express includes CDH and a version of Cloudera _____ lacking enterprise features such as rolling upgrades and backup/disaster recovery.
Option A:	Enterprise
Option B:	Express
Option C:	Standard
Option D:	Manager
Q15.	_____ technique allows us to filter streams so elements that belong to a particular set are allowed through, while most nonmembers are deleted
Option A:	Bloom Filters
Option B:	DGIM
Option C:	FM
Option D:	Window filter
Q16.	_____ is a function that assigns a real number to each page in the Web. The intent is that the higher the number of a page, the more “important” it is.
Option A:	Term Spam
Option B:	Page Rank
Option C:	Window size
Option D:	Teleport
Q17.	The techniques for artificially increasing the PageRank of a page are collectively called
Option A:	Link Spam
Option B:	Term Spam
Option C:	Teleport

Option D:	Page Rank
Q18.	A collection of pages whose purpose is to increase the PageRank of a certain page or pages is called as a _____.
Option A:	Link Spam
Option B:	Term Spam
Option C:	Teleport
Option D:	Spam Farm
Q19.	_____ is topic-sensitive PageRank, where the “topic” is a set of pages believed to be trustworthy (not spam).
Option A:	Link Spam
Option B:	Term Spam
Option C:	Trust Rank
Option D:	Spam Farm
Q20.	This hubs-and-authorities algorithm, sometimes called HITS
Option A:	HITS
Option B:	DGIM
Option C:	FM
Option D:	Page Rank
Q21.	Certain pages are valuable because they provide information about a topic. These pages are called _____
Option A:	Authorities
Option B:	Hubs
Option C:	Page Rank
Option D:	HITS
Q22.	Some pages are valuable not because they provide information about any topic, but because they tell you where to go to find out about that topic. These pages are called _____.
Option A:	Authorities
Option B:	hubs
Option C:	Page Rank
Option D:	HITS
Q23.	_____ is the process of gathering and analyzing data from social networks such as Facebook, Instagram, LinkedIn and Twitter.
Option A:	Web mining
Option B:	Text mining

Option C:	Multimedia mining
Option D:	Social media analytics
Q24.	_____ a form of data mining operating on data stored in multiple database tables.
Option A:	MRDM
Option B:	DREM
Option C:	RETM
Option D:	VGTM
Q25.	_____ is a subfield of data mining which is used to find interesting information of implicit knowledge from multimedia databases
Option A:	Text mining
Option B:	Multimedia mining
Option C:	Pattern mining
Option D:	Class mining