BRAINWARE UNIVERSITY

Term End Examination 2022
Programme – MCA-2020/MCA-2021
Course Name – Big Data Analysis
Course Code - MCA305A
(Semester III)

Full Marks : 60 **Time : 2:30 Hours**

[The figure in the margin indicates full marks. Candidates are required to give their answers in their own words as far as practicable.]

Group-A
(Multiple Choice Type Question) 1 x 15=15

1. *Choose the correct alternative from the following :*

(i) Report the identifiers in HiveQL are
a) case sensitive
b) case insensitive
c) sometimes case sensitive
d) Depends on the Hadoop environment

(ii) Justfy the correct option for the blank: Most of the _____data generarted on twitter
a) Structured data
b) Unstructured data
c) Semi structured data
d) None of these

(iii) Define the language that Hadoop written in.
a) Java (software platform)
b) Perl
c) Java (programming language)
d) Lua (programming language)

(iv) State the main components of big data.
a) HDFS
b) YARN
c) Map Reduce
d) All of these

(v) Define the platform in which Hadoop runs on.
a) Bare metal
b) Debian
c) Cross-platform
d) Unix-like

(vi) Identify the object which maps input key/value pairs to a set of intermediate key/value pairs.
a) Mapper
b) Reducer
c) Both Mapper and Reducer
d) None of the mentioned

(vii) Report the situation when you consider an association rule interesting?
a) If it only satisfies min_support
b) If it only satisfies min_confidence
c) If it satisfies both min_support and min_confidence
d) There are other measures to check so

(viii) Observe the expression of Confidence (A -> B) and select the correct answer.

a) Support(AUB) / Support
b) Support(AUB) / Support (B)
c) Support(AUB) / Support (A)
d) None of mentioned

(ix) Judge the requirement of K-means clustering?

a) defined distance metric
b) number of clusters
c) initial guess as to cluster centroids
d) all of the mentioned

(x) Determine the algorithm that is most sensitive to outliers?

a) K-means clustering algorithm
b) K-medians clustering algorithm
c) K-modes clustering algorithm
d) K-medoids clustering algorithm

(xi) Write the algorithm that does not require a dendrogram?

a) K-means
b) Agglomerative
c) Divisible
d) All of the mentioned

(xii) Calcute the different features of Big Data Analytics.

a) Open source
b) Data recovery
c) Scalibility
d) All of these

(xiii) Select the correct statements about sampling.

a) Sampling reduces the amount of data fed to a subsequent data mining algorithm.
b) Sampling reduces the diversity of the data stream.
c) Sampling aims to keep statistical properties of the data intact.
d) Both a and c

(xiv) Choose The correct answer for the blank: Most NoSQL databases support automatic _____ meaning that you get high availability and disaster recovery.

a) Processing
b) Scability
c) Replication
d) All of these

(xv) Decide the types of nosql databases.

a) Graph & Column-oriented databases.
b) Document database
c) Key-value stores
d) All of the above

Group-B

(Short Answer Type Questions) 3 x 5=15

2. Illustrate the features of NoSQL. (3)
3. Illustrate unsupervised learning method with example. (3)
4. Explain likelihood and evidence in Naive Bayes. (3)
5. What is sentiment analysis with its challenges. (3)
6. Illustrate the various types of data in big data. (3)

OR

Discuss the applications of Association rules. (3)

Group-C

(Long Answer Type Questions) 5 x 6=30

7. Illustrate the workflow of YARN in detail. (5)
8. Illustrate decision trees classification. (5)
9. Explain in detail about HDFS. (5)
10. Explain content-based filtering in detail. (5)
11. Evaluate why we use NoSQL and how it deffers from Relational database? (5)
12. Define the characteristics of Big Data Applications. (5)

OR

Describe the components of Hadoop architecture in details. (5)