Avinashilingam Institute for Home Science and Higher Education for Women
(Deemed to be University) Coimbatore-641043.

Master's Degree Examination – November 2018

III Semester

Class : II PG **Time: 3 hours**
Major : Computer Science **Max. Marks: 60**

17MCSC14 INTRODUCTION TO DATA SCIENCE

Part A **10 x 1/2 = 5**

Choose the correct answer

1. Which of the following is most important language for Data Science ?

   a. Java b. Ruby c. R d. Python

2. Which of the following is performed by Data Scientist?

   a. Define the question b. Create reproducible code c. Challenge results d. All of the above

3. What are the three V's of Big data?

   a. Volume b. Velocity c. Variety d. All the above

4. What are the different features of Big Data analytics?

   a. Open Source b.Scalability c. Data Recovery d. All the above

5. The most convenient way to use R is at a graphics workstation running a _____ system.

   a. Windowing b. Running c. Interfacing d. All of the above.

6. What would be the result of following R code?

   >x->1
   >print(x)
   a.1 b.2 c.3 d.4

7. _____ is a platform for constructing data flows for extract, transform, and load (ETL) processing and analysis of large datasets.

   a. Pig Latin b. Oozie c. Pig d. Hive

8. A _____ node acts as the Slave and is responsible for executing a Task assigned to it by the JobTracker.

   a. MapReduce b. Mapper c. TaskTracker d. JobTracker

9. IBM and _____ have announced a major initiative to use Hadoop to support university courses in distributed computer programming.

   a. Google Latitude b.Android c. Google Variations d. Google

10. Hadoop is a framework that works with a variety of related tools. Common cohorts include:
    a. MapReduce, Hive and HBase b. MapReduce, MySQL and Google Apps
    c. MapReduce, Hummer and Iguana d. MapReduce, Heron and Trumpet

Part B 5 x 4 = 20

Answer ALL questions

Each answer should not exceed 200 words or one page

11.a. Briefly write about the importance of data

(Or)

11.b. Write down the roles of Big data ecosystem.

12.a. Write short notes on data preparation in analytics life cycle

(Or)

12. b. What is the need of communicate results in analytics life cycle? Explain.

13.a. How data import and export is done in R?

(Or)

13.b. Write a note on dirty data and how it affects data analysis? Justify.

14.a. Write about Apache Pig

(Or)

14.b. What are ordered aggregates in SQl?

15.a. What are the characteristics of Big data? Explain

(Or)

15.b. How to get data in Hadoop? Explain.

Part C 5 x 7 = 35

Answer ALL questions

Each answer should not exceed 600 words or three pages

16.a. Write about the data structures of big data with suitable examples.

(Or)

16.b. Compare emerging Big data Eco system and a new approach Analytics.

17.a. Explain in detail about data analytics life cycle

(Or)

17.b. How model planning and model building is so important in an analytics life cycle. Justify.

18.a. Explain how exploratory data analysis is done in R.

(Or)

18.b. Compare and contrast data exploration and presentation.

19.a. Explain the Hadoop ecosystem with neat diagram.

(Or)

19.b. What are windows functions and user defined functions in SQL. Explain in detail.

20.a. Write down how big data is viewed from business perspective.

(Or)

20.b. Write about Hadoop in detail.