H1064 **Pages 3**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Scheme for Valuation/Answer Key

Scheme of evaluation (marks in brackets) and answers of problems/key

EIGHTH SEMESTER B.TECH DEGREE EXAMINATION, MAY 2019

Course Code: CS466

Course Name: DATA SCIENCE

Max. Marks: 100 Duration: 3 Hours

PART A

Answer all questions, each carries 4 marks. Marks

1 Different roles of in data science project (project sponsor, client, data scientist, data architect, operations) (4)

2 Brief note on accessing text files using R programming languages - 2 mark (4)

Syntax and functions for data access(read. table() command) – 2 mark

3 4 distance measures (Euclidean , Manhattan, hamming. Cosine) (4)

4 X<- array(1:20,dim=c(2,5)) (4)

For printing

X

5 Explanation(2 mark) + syntax (2 marks) (4)

6 Answer-2 marks (4)

2 rows

2 columns

Position 1

Explanation of the function – 2marks

7 Any four advantages (4)

8 Name nodes and data nodes (4)

9 To produce reproducible work; inclusion of R code and results inside documents. (4)

10 matplot() or pairs() (4)

PART B

Answer any two full questions, each carries 9 marks.

11 a) Stages of data base project with figure(defining goals, data collection and management, modelling, model evaluation critique, presentation and documentation, model deployment and maintenance) 6 mark (9)

Example – 3 mark

12 a) List problems which involve machine learning as technique in solution(classification, scoring, clustering , association rules) – 4 mark (9)

Problem to method mapping (5)

13 a) Logistic regression (basics , no need of code)— 3 mark (3)

b) Linear regression -- 2 marks (6)

Building the models and making predictions -- 4 mark

PART C

Answer any two full questions, each carries 9 marks.

14 a) Data frame definition+ syntax 2 marks (6)

attach() –1 mark

detach – 1 mark

search() –2 mark

b) X<-lm(*formula, data=data. frame*) with example (3)

15 a) Any 4 probabilistic distribution functions (4)

b) Any formula with explanation 3 marks+ 2 examples 2 marks (5)

Or

Explain generally Linear regression and ANOVA etc

16 a) Collaborative Filtering - User-Item (4 Marks) (9)

Use of Euclidean distance to explain collaborative filtering with any example like Movie rating(3 Marks)

Python code or R code or pseudo code(2 Marks)

PART D

Answer any two full questions, each carries 12 marks.

17 a) Figure(2)+ explanation(2 marks) (4)

b) (8)

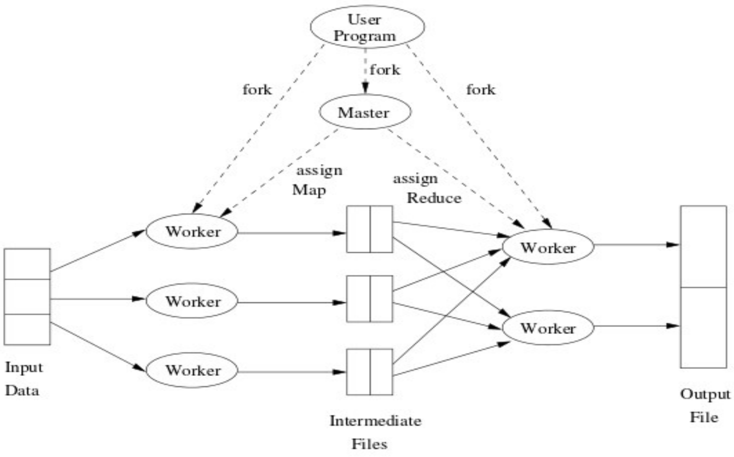

Figure 2.3: Overview of the execution of a MapReduce program

18 a) HDFS Failover architecture (6)

Or

(The question is about failures in hadoop map reduce, but students can refer to any one of the books in the syllabus which explains only failures in map reduce So marks can be awarded to failures in map reduce also)

b) mfrow(3,2)- 3 rows and 2 colums (allot in row order) (6)

mfcol(3,2)- 3 columns and 2 rows(allot in column order)

19 a) To sponsor (4 marks) (12)

To end users(4 marks)

To data scientists(4 marks)

****