

MANIPAL UNIVERSITY
SCHOOL OF INFORMATION SCIENCES

SECOND SEMESTER MASTER OF ENGINEERING – ME (COMPUTING TECHNOLOGIES &
VIRTUALIZATION)
DEGREE EXAMINATION – APRIL / MAY 2016

SUBJECT: VIR 610.6 (ELECTIVE 2) – BIG DATA ANALYTICS

Friday, May 6, 2016

Time: 10.00 – 13.00 Hrs.

Max. Marks: 100

1. In the data acquisition phase, especially in the data pre-processing stage of a Big Data system, analyze the following steps:
 - (a) data integration.
 - (b) data cleansing.
 - (c) redundancy removal.

(3 + 3 + 4 = 10 marks)
2. In the context of healthcare solutions, identify the key challenges with respect to velocity, volume, variety, value, and veracity characterization of big data systems.

(2 x 5 = 10 marks)
3. A. Identify at least three key differences between the following two database technologies.

(5 marks)

B. For each of these technologies mention one specific application context where it shines over the others, that is, it improves the solution.

 - (a) document-based databases.
 - (b) column-based data stores.

(5 marks)
4. In the context of Big Data systems, analyze the differences among batch processing and stream processing systems with an educated example for each.

(5 + 5 = 10 marks)
5. A. State the different metadata maintained by GFS during its operations.

(5 marks)

B. With an educated example, state the role of each component during the execution of a client's request.

(5 marks)

6. A. What is the essential nature of the MapReduce programming model?
(5 marks)
B. Define the meaning of map and reduce operations with a simple example.
(5 marks)
7. Show how Google's MapReduce system handles
(a) Master failure
(b) Worker failure.
(5 + 5 = 10 marks)
8. Consider the problem of searching for documents (which may be web pages) that match a given criteria. Show how MapReduce techniques may be used to structure the solution. Show the different components of such a system in a neat diagram.
(4 + 6 = 10 marks)
9. Describe the technical aspects of the data model of Google's Bigtable implementation. Why does Bigtable depend on a distributed lock service (Chubby) for its operation?
(5 + 5 = 10 marks)
10. A. What is storytelling in Visualization?
B. What is Web Scrapping?
C. What are the different ways web scrapping can be done?
(4 + 3 + 3 = 10 marks)
