# Big Data Analytics Question Bank

Module 1

1. Describe the various components of HDFS with a neat diagram. [10]
2. Write a note on the following:
   a>HDFS Block replication b>HDFS safe mode c>Rack awareness [4+3+3]
3. Describe basic steps in MapReduce parallel data flow with neat diagram.[10]
4. With a neat diagram discuss the following : a. HDFS Namenode high availability design
   b. HDFS namenode federation [5+5]
5. What are HDFS snapshots? List out the features they offer. [5]
6. List and explain file manipulation commands in HDFS. [10]
7. Write and explain simple mapper and reducer script to search particular word. (book 1 -page no 102 ) [10]
8. Is Mapreduce fault tolerant? Justify your answer (5)
9. What is speculative execution?

2.

1. List important properties of MapReduce. [5]
2. Write note on Combiner step, placement of mapper and reducers with suitable diagrams. [10]
3. Write and Explain WordCount program in java. Also write steps to execute this program in Hadoop framework .[10]
4. Write and explain steps to execute HDFS C application example to manipulate HDFS files and file system. [10]
5. Explain use of Pipes library with example. [10]
6. Explain MapReduce chaining with example. [10]

Module 2

1. What is Apache Pig? What are the modes of operation (4)
2. What id Apache Hive? List its features
3. Write the commands to start Hive, create a table, Show a table and drop a table in Hive (2)
4. Explain the process of Data import and export in Sqoop with neat diagrams (8)
5. Compare the Version 1 and Version 2 of Sqoop (5)
6. What is Apache Flume? Explain the components of a Flume agent with neat diagram (8)
7. Describe a Flume pipeline and Flume consolidation network with diagrams (6)
8. Explain in detail how Hadoop workflows are managed with Apache Oozie (10)
9. Describe the Hbase Data Model and list its important features (8)
10. Describe the structure of YARN application (8)
11. With a neat diagram explain the Hadoop version 2 ecosystem (10)
12. Write a brief note on the following (8)
    - Apache REEF,
    - Hamster
    - Apache Flink
    - Apache Slider

13. What are the features available on Ambari dashboard view of Hadoop cluster (6)
14. Discuss the features provided in the Services and Host view of Ambari (8)
15. What are the three options available in the Admin View of Ambari (4)
16. How do you set the container memory in YARN (4)
17. What is a HDFS balancer tool? Explain in detail (8)

Module 3

1. Define business intelligence and explain the BIDM cycle.
2. What is the role of BI in decision making? List the skillset of a BI specialist
3. Differentiate between operational and strategic decisions with example
4. Explain any five business intelligence applications
5. List the requirements of a good data warehouse
6. Differentiate between a functional data mart and Enterprise Data warehouse
7. Describe the data warehouse architecture?
8. What is ETL? List steps of ETL process.
9. What is the concept of data transformation in a data warehouse?
10. Expand OLAP in data warehouse technology and Explain?
11. What is data mining? Explain various data mining tasks.
12. Why is Data cleaning and preparation important in Data mining?
13. Discuss about the outputs produced by the Data mining process?
14. What is confusion matrix?
15. What are unsupervised and supervised learning techniques
16. List the tools and platforms used for data mining
17. Explain the CRISP-DM Data mining cycle
18. List some of the common myths of data mining
19. List some of the common mistakes made in Data mining
20. What is data visualization? How do you achieve excellence in data visualization
21. What are the different types of charts used for data visualization
22. List out some important considerations/Tips for better data visualization

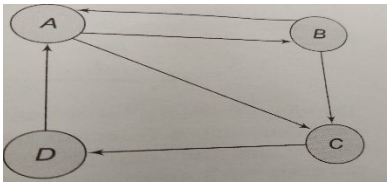
Module -4

1. What is a decision tree? Give the pseudocode and explain the steps in building a decision tree
2. What is pruning? Discuss the two approaches of pruning
3. Explain correlations among data elements and define correlation coefficient
4. What is Logistic Regression?
5. Discuss the advantages and disadvantages of Regression
6. Describe the design principles of ANN and list its business applications [5+5]
7. Briefly describe the ANN architecture for Neural Networks
8. What are steps required to build ANN? What are advantages and disadvantages of ANN. [10]
9. How do you represent an association rule? Explain with example. [10]
10. What is a cluster? Write the generic pseudo code for clustering. [05]
11. List the advantages and disadvantages of K – Means clustering. [05]
12. Describe three business applications in industry where cluster analysis will be used.
13. Describe K-means clustering with an example. [10]

14. What is Association rule mining? List its business applications [6]
15. How are association rules reperesented
16. What are frequent itemsets? How does Apriori algorithm works? [10]
17. Solve the example and exercise problem from decision tree [Page 77 and 87]
18. Solve the example and exercise problem from Association rule mining [Page 126 and 131]

Module 5

23. With a neat diagram explain the text mining process (5)
24. What is term document matrix? Explain with example (5)
25. What are the considerations to be taken into account while creating a TDM (5)
26. What are the various techniques to mine the TDM (5)
27. Compare text mining and Data mining (6)
28. What are the best practices to be applied during the text mining process? (4)
29. What is Naïve Bayes technique? Explain the model with a simple classification example (10)
30. How is Naïve Bayes applied to text mining explain with example (10)
31. What are the advantages and disadvantages of Naïve-Bayes (3)
32. What is SVM? Explain the SVM model (8)
33. Describe the kernel method (6)
34. What are the advantages and disadvantages of SVM's? (5)
35. What is Web mining? List the characteristics of optimized websites (5)
36. Describe the three different types of web mining (10)
37. Write a note on the following: a. Web mining algorithm b. click stream analysis (6)
38. What is Social Network Analysis (SNA)? List its applications (6)
39. Describe the Ring and Hub network topology (6)
40. Describe the Influence Flow Model to compute the importance of a node and compute the rank values for the nodes of the following network (10)




41. Write a note on the following a. Finding subnetworks b. Pagerank (8)
42. How is social Network Analysis different from other Data mining techniques?(5)
43. What are the challenges faced in analyzing SNA (4)