UNIT 3 Question Bank

1. What is Apache Spark? Explain some key features of Spark.

2. What are the benefits of Spark over MapReduce? Please explain.

3. Can you use Spark to access and analyse data stored in Cassandra databases?

4. What are the languages supported by Apache Spark for developing big data applications?

5. Explain about the different cluster managers in Apache Spark.

6. Explain about the major libraries that constitute the Spark ecosystem?

7. What is Executor Memory in a Spark application? Please explain.

8. What are the various data sources available in SparkSQL?

9. What do you understand by Pair RDD?

10. Please explain the difference between Hadoop MapReduce and Spark.

11. What is RDD in Spark and how it can be created in different ways?

12.Explain each way with example code snippet.

13. What are the different modes to run Spark? What Spark mode is the best industry practice and why? Please explain.

14. Please explain the Spark architecture with each and every internal component, such as Spark Driver, Worker Node and Spark Executor.

15. How Spark RDD can be partitioned in Spark Cluster? What do you mean by the term 'Resilient' in Spark? Please explain.

16. What is Dataframe and Dataset in Spark? When and why do we need to create Dataframe and Dataset in Spark? Please explain with some examples.

17. Explain about the different cluster managers in Apache Spark.

18. What are the features present in the Spark architecture that enable fast computations and usages of expressive programming model? (LO 5.1)

19. Describe the functions of Spark SQL, Spark Streaming and GraphX? (LO 5.1)

20. How do Spark and Python provide a powerful Big Data analysis tool? (LO 5.2)

21. How does DataFrame create from JSON datasets and Hive tables? (LO 5.2)

22. What are the aggregation commands provisioned in Spark SQL? (LO 5.2)

23. How do NumPy, SciPy and Pandas Python libraries provision for advanced functions for analytics, and create an integrated development environment (IDE)? (LO 5.2)

24. How does the Spark Resilient Distributed Dataset (RDD) programming collect the objects?(LO 5.3)

25. Explain method of creation of RDDs using the transform and action commands. (LO 5.3)

26. Explain the need for computing the time complexities of algorithm.

27. Explain wall clock time and communication cost complexity.

28. What is the communication cost of a Naïve's algorithm.

29. How is parallelism achieved in Naive's algorithm.

30. How do you compute the time complexity of a 2 phase map reduce job.