COMP9313 20T2 DengDeng review questions

Name and section: _____

1 Multiple choice

1. Which of the "Vs" listed below were the three first Vs proposed to characterize big data?
    A. Value, Volume, Velocity
    B. Volume, Velocity, Viability
    C. Validity, Variability, Volume
    D. Velocity, Volume, Variety

2. Which of the options listed below include activities/tasks corresponding to the Data Management phase in Big Data Processes?
    A. Cleaning, Modelling and Analysis
    B. Aggregation, Recording, Interpretation
    C. Acquisition, Analysis and Interpretation
    D. Aggregation, Representation, Integration

3. Which of the following options is true regarding Apache Spark?
    A. Provides better support for iterative algorithms and keeps more data in memory
    B. Keeps more data in memory, but totally lacks support for iterative algorithms
    C. Does not support interactive shells
    D. Is a modified version of Apache Hadoop

4. Which of the following options is true about Apache Hadoop?
    A. Programmers need to explicitly program for and handle failures and data loss
    B. Apache Hadoop does not support job coordination
    C. Apache Hadoop provides a fault-tolerant data storage
    D. Apache Hadoop is a modified version of Apache Spark

5. The Big Data Process involves a number of phases and activities. The success of a big data project can be guaranteed only when such process is executed end-to-end and sequentially (no iterations involved) from acquisition and recording, to interpretation.
    A. True
    B. False

6. Which of the following statements is true regarding shuffle and sort in MapReduce?
    A. Shuffle and sort is in charge of aggregating the outputs produced by Mappers based on the produced value of each $< key, value >$ pair
    B. The logic for shuffling and sorting must be defined by the programmer
    C. Shuffle and sort is automatically handled by the framework
    D. None of the above

7. Which of the following statements is true about MapReduce?

A. Programmers must always be aware of where map and reduce is running in the cluster, as otherwise the solution will not be correct

B. Programmers must keep track of which key is processed by each reducer

C. Programmers must provide the logic (implementation) for map tasks

D. None of the above

8. Lazy evaluation in Spark helps to:

A. Perform an operation on an RDD right when a transformation is requested by the user, but only when using the Spark interpreter

B. Postpone the execution of an Action to a later time only when a transformation operation is requested

C. Perform optimizations (by Spark) that can improve the overall performance of the program

D. None of the above

9. Transformations in Spark allows for the:

A. Execution of an operation that results in the creation of a new RDD

B. Modification of the data stored in an RDD (in-situ)

C. Transformation of an RDD into a value (val) that can be handled by Scala

D. Transformation of an RDD into a new Action to be lazily evaluated

10. Which of the following statements is true about Actions in Spark?

A. Actions are operations that return a value

B. Actions must never be called before all Transformations are requested in a Spark program, regardless of whether the Action depends on the Transformations in a program

C. Actions are always lazily evaluated and triggered by a Transformation request

D. All Actions must be provided with an anonymous function as an argument in order to be executed

11. Besides HDFS, Spark can also interact with other storage systems such as S3, local file system and HBase:

A. True

B. False

12. Which of the following statements is true about RDD persistence?

A. RDD persistence causes the RDD to be always stored in full on disk

B. You must always use RDD persistence to guarantee the correctness of your program

C. Without RDD persistence, node failures will cause to completely lose RDDs without any possibility of recomputing it

D. RDD persistence helps in storing intermediate results (RDDs) and avoiding re-computation of the RDD

13. Which of the scenarios below provide a strong case to choose HBase over a traditional RDMS (assuming data can be stored in rows/columns)?

A. Source data and business requirements may frequently change causing frequent addition / deletion of rows in the database

B. Schema of the database is well known in advance and it is unlikely to change over time

C. Source data and business requirements may frequently change causing frequent addition / deletion of table attributes in the database

D. All of the above

14. Which of the statements below can be considered an integration task in the Cyber security scenario discussed in the lecture?
    A. The retrieval and storage of security vulnerability information from multiple sources available on the web
    B. The enrichment of collected information with security-related information to support complex queries
    C. The removal of potential noise found in the collected security vulnerability information
    D. None of the above

15. Which of the following Big Data technologies can be used to train ML models to support natural language interactions (e.g., using bots)?
    A. HBase
    B. Plain HDFS
    C. Spark MLlib
    D. Map Reduce

16. Consider the characteristic of having a very-high dimensional space in the context of finding similar items in a dataset. Which of Vs (listed below) of Big Data is more closely related to this characteristic?
    A. Variety
    B. Visibility
    C. Velocity
    D. Volume

17. The problem of not being able to access relevant events that are needed to perform process mining tasks can be considered as related to the Visibility dimension of Big Data.
    A. True
    B. False

18. which statement is true about Spark:
    A. Spark can only be runned by using Hadoop YARN as the Cluster
    B. Spark can only be runned on HDFS
    C. Spark is not a modified version of Hadoop
    D. None of above

19. which statement is true about Spark:
    A. RDD Persist can be considered as a execution trigger for a RDD in Spark
    B. By using function cache() in Spark java, RDD can only be stored in the memory
    C. In Spark java fiction foreach can be views as a action
    D. None of above

20. which statement is true about MapReduce:
    A. Stage of shuffle is proceed after reducer
    B. In Hadoop 1.x, YARN can be used for managing resourses
    C. Compared with RDBMS, Hadoop can take unstructured data as input
    D. We can track which intermediate key a particular reducer is processing

21. In regard to Big Data security challenges, relying on techniques to anonymize the data can guarantee security and privacy
    A. True
    B. False

22. Why is it important to have monitoring mechanisms in place when dealing with Big Data Security?
    A. Because it can improve the quality of data
    B. Because it provides a strong authentication mechanism for big data frameworks
    C. Because no system is 100% secure
    D. All of the above

23. In Hadoop Distributed File System (HDFS), the NameNode is responsible for
    A. Storing the data records
    B. Storing Metadata about the data records
    C. Submitting MapReduce jobs

24. Which of the following statements is true regarding shuffle and sort in MapReduce
    A. Shuffle and sort is in charge of the final aggregation of the outputs produced by Mappers based on the produced value of each ¡key, value¿ pair.
    B. The logic for shuffling and sorting must be defined by the programmer
    C. Shuffle and sort is automatically handled by the framework
    D. None of the above

25. Which of the following statements is true about MapReduce
    A. Programmers must always be aware of where map and reduce is running in the cluster, as otherwise the solution will not be correct
    B. Programmers must provide the logic (implementation) for map tasks
    C. Programmers must keep track of which key is processed by each reducer
    D. None of the above

26. What is NOT true about the Map fucntion in MapReduce?
    A. the application of the function happens in isolation
    B. Map tranforms the input into key-value pairs to process
    C. Map aggregate the list of values for each key

27. Besides HDFS, Spark can also interact with other storage systems such as S3, local file system and HBase
    A. True
    B. False

28. Which of the following statements is true about RDD persistence?
    A. RDD persistence causes the RDD to be always stored in full on disk
    B. Without RDD persistence, node failures will cause to completely lose RDDs without any possibility of recomputing it
    C. RDD persistence helps in storing intermediate results (RDDs) and avoiding re-computation of the RDD

2 Short answer

1. In the original Hadoop Framework, we usually use HDFS to storage data and _____to process data and use _____to manage resources.
2. In the HDFS architecture, relicas information are stored in _____.
3. The biggest limitation of MapReduce is that it read and write all the data to _____, while Spark keep them in _____.
4. The cluster is setup when a _____object is created in the programming of Spark.
5. Spark _____is assigned by Cluster manager and is responsible for running computations and accesses data storage.
6. In spark, the reason why lazy evaluation can be realized is that RDD transformation information are being stored as _____by Spark Driver.