Home Assignment – Big Data Fundamentals

Objective Questions

Ques1: The primary Machine Learning API for Spark is now the _____ based API

1 **DataFrame**

2 Dataset

3 RDD

4 All of the above

Ans. DataFrame

Ques2: _____ is a component on top of Spark Core.

1 Spark Streaming

2 Spark SQL

3 RDDs

4 All of the above

Ans. Spark SQL

Ques3: Given a dataframe df, select the api/function that returns its number of rows:

1 df.take('all')

2 df.collect()

3 df.count()

4 df.numRows()

Ans. df.count()

Ques4: Given a DataFrame df that includes a number of columns among which a column named quantity and a column named price, complete the code below such that it will create a DataFrame including all the original columns and a new column revenue defined as quantity*price (Scala Lang) :

1 df.withColumnRenamed("revenue", expr("quantity*price"))

2 df.withColumn(revenue, expr("quantity*price"))

3 df.withColumn("revenue", expr("quantity*price"))

4 df.withColumn(expr("quantity*price"), "revenue")

Ans. df.withColumn("revenue", expr("quantity*price"))

Ques5: Which of the following is true for RDD?

1 We can operate Spark RDDs in parallel with a low-level API

2 RDDs are similar to the table in a relational database

3 It allows processing of a large amount of structured data

4 It has built-in optimization engine

Ans. We can operate Spark RDDs in parallel with a low-level API

Ques6: SparkSQL translates commands into codes. These codes are processed by
1 Driver nodes
2 Executor Nodes
3 Cluster Manager
4 None of the above

Ans. Executor Nodes

Ques7: The shortcomings of Hadoop MapReduce was overcome by Spark RDD by
1 Lazy-evaluation
2 DAG
3 In-memory processing
4 All of the above

Ans. All of the above

Ques8: Which of the following is a distributed graph processing framework on top of Spark?
1 Spark Streaming
2 MLlib
3 GraphX
4 All of the above

Ans. GraphX

Ques9: Which of the following is the reason for Spark being faster than MapReduce while execution time?
1 It supports different programming languages like Scala, Python, R, and Java. 2
RDDs
3 DAG execution engine and in-memory computation (RAM based)
4 All of the above

Ans. DAG execution engine and in-memory computation (RAM based)

Ques10: Each kafka partition has one server which acts as the _____
1 leader
2 followers
3 staters
4 All of the mentioned

Ans. leader

Ques11: Which all are the elements of Kafka?
1 Topic
2 Producer
3 Consumer
4 All of these

Ans. All of these

Ques12: What of the following is true w.r.t consumers in Kafka?
1 If all consumer instances have the same consumer set, then this works like a conventional queue adjusting load over the consumers
2 If all customer instances have dissimilar consumer groups, then this works like a publish-subscribe and all messages are transmitted to all the consumers
3 Both A and B
4 None

Ans. Both A and B

Ques13: Kafka maintains feeds of messages in categories called
1 Topics
2 Chunks
3 Domains
4 Messages

Ans. Topics

Ques14: Kafka only provides _____ order over messages within a partition
1 Partial
2 Total
3 30%
4 None of the mentioned

Ans. Total

Ques15: Which all are Kafka key capabilities?
1 Publish and subscribe to streams of records, similar to a message queue or enterprise messaging system
2 Store streams of records in a fault-tolerant durable way
3 Process streams of records as they occur
4 All of these

Ans. All of these

Ques16: The kafka-topics CLI needs to connect to.?
1 Zookeeper
2 Broker
3 Topic
4 None of the above

Ans. Zookeeper

Ques 17: In Kafka records are published to:
1 Table
2 Subject
3 Topic
4 None of the above

Ans. Topic

Ques18: A Kafka record is uniquely identified within the Partition by its _____?
1 Timestamp
2 Broker
3 Primary Key
4 Offset

Ans. Offset

Ques19: Suppose a Producer has written a message to Kafka. That message can be changed.
1 Anytime, by any Producer
2 Only by the Producer who sent it to Kafka
3 Only to change its metadata
4 Never

Ans. Never