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

Ques2: ____________ is a component on top of Spark Core.
1 Spark Streaming
2 Spark SQL
3 RDDs
4 All of the above

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()

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")

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

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

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

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

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

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

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

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

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

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

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

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

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

Coding Assignment

For below mentioned exercise, share relevant code and snapshots

A. Spark Batch -

   Read attached json (demographic_info.json) into a Spark Dataframe and carry out following -

   1. Use following apis over Dataframe
      select to show columns (name, age, gender, isActive, balance, company, eyeColor, email, phone)
      filter to show records with isActive as true
   2. Show top 2 male and female with maximum balance
   3. Add a column Age_group with classifications as Teenager (13-19 years), Young (20-40 years), Old (>40 years)
   4. Create temp table view over initially read json Dataframe and run sql queries for same requirements given above.
   5. Convert above selected column dataframe into an RDD and save it into a text file.

B. Kafka –

   1. Create a topic with 3 partitions
   2. Create a producer writing data to above created kafka topic with following considerations –
      a. messages should be read from a file line by line (use any file from your side with limited content)
      b. message should be produced on kafka topic in (key, Value) format where key is timestamp+index and value is actual message

C. Structured Streaming -

   1. Create a Spark streaming job reading data from above created kafka topic with following considerations –
      a. read from beginning
      b. calculate word count
      c. print word count to console