1. "What does the code below print:

var pokemon = List(""Pikachu"", ""Raichu"", ""Psyduck"", ""Pidgey"", ""Blastoise"",""Bulbasaur"")

println {

pokemon.filter { (pokedex: String) =>

pokedex.take(1) == ""P""

}.reduceLeft { (a: String, b: String) =>

s""$a $b""

}

}"

1. Pikachu Psyduck Pidgey
2. Pikachu
3. Raichu
4. Throws an error

Answer: a

1. Which of the following is a wide transformation?
2. Map
3. Filter
4. groupByKey
5. flatMap

Answer: c

1. Which of the following is not a DStream output operation?
2. SaveAsTextFiles
3. ForeachRDD
4. Print
5. ReduceByKeyAndWindow

Answer: d

1. Which of the following action operations does not return the result the driver?
2. collect()
3. take(n)
4. first()
5. foreach()

Answer: d

1. Which of the following statements is TRUE regarding stateless transformation?
2. Computation of current results requires results from privious operation
3. Windowed operations are Stateless transformations
4. Each batch of data is computes independently
5. updateStateByKey() is a type of Stateless transformation

Answer: c

1. Which of the following is not true for Catalyst Optimizer?
2. Catalyst optimizer makes use of pattern matching feature.
3. Catalyst contains the tree and the set of rules to manipulate the tree.
4. There are no specific libraries to process relational queries.
5. There are different rule sets which handle different phases of query.

Answer: c

1. What are the advantages of using larger batch sizes when processing Spark SQL tables/dataframes.
2. Improve memory utilization
3. Improve compression efficiency
4. Less 'Out Of Memory' exceptions when caching data
5. Better Garbage Collection

Answer: a and b

1. Assume that you have an RDD of tuples containing (student\_name, subject, marks) of type (string, string, int). If you want to get the maximum marks for each student from all subjects such as (student\_name, max\_marks) where max\_marks is the student's highest score from ll subjects, what is the most efficient transformation method you can use?
2. mapPartitions
3. aggregateByKey
4. reduceByKey
5. join

Answer: b

1. Suppose you have a small lookup table whose data is static. You want to perform a distributed join on the this lookup data with another fact table with is huge and distributed across several executor nodes in your cluster. For performance reasons, you don't to send a copy of the lookup tables data to every task running on the executor. Instaed, you want to limit only one copy for all tasks running on each node. Which technique do you use to achive this.
2. Make Lookup table as an Accumulator
3. Make lookup table as a Broadcast variable
4. Persist the look table dataframe
5. Convert lookup table into a Dataframe.

Answer: b

1. Which of the following are the valid considerations in tuning memory usage:
2. the amount of memory used by your objects
3. the cost of accessing those objects
4. the number of objects to serialized
5. the overhead of garbage collection
6. the serialization format of your data

Answer: a, b and d

1. FlatMap transforms an RDD of length N into another RDD of length M. which of the following is true for N and M. Which of the following is true?
2. N >= M
3. N == M
4. N <= M
5. N > M

Answer: c

1. Assume that you have an RDD of integers and you want to output top 6 biggest integers from the RDD. Which of the following code snippets can do this?
2. rdd.take(6, key=lambda x: -x, -1)
3. rdd.takeSorted(6, key=lambda x: x)
4. rdd.takeOrdered(6, key=lambda x: x)
5. rdd.takeOrdered(6, key=lambda x: -x)

Answer: d

1. Which property control the number of executors requested to run your spark application on the cluster
2. –executors
3. --spark-executors
4. --num-executors
5. –total-executors

Answer: c

1. Which of the following statements about the Workspace of Azure Databricks concept is false
2. It has a notebook, library, and dashboard storage.
3. It oversees data ETL activities.
4. It is the Azure Databricks root folder.
5. It assists in Stream Processing Jobs

Answer: b

1. Databricks' authentication and authorization can be controlled for
2. Person, Group
3. List of access controls
4. Group, User, and Access Control List
5. Access Control List, Group

Answer: c

1. What is a lambda architecture and what does it try to solve
2. An architecture that defines a data processing pipeline whereby microservices act as compute resources for efficient large-scale data processing
3. An architecture that employs the latest Scala runtimes in one or more Databricks clusters to provide the most efficient data processing platform available today
4. An architecture that splits incoming data into two paths - a batch path and a streaming path. This architecture helps address the need to provide real-time processing in addition to slower batch computations.
5. Design Patterns for optimizing Spark Jobs

Answer: c

1. You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table. Which output mode should you use
2. Update
3. Complete
4. Merge
5. Append

Answer: d

1. What happens to Databricks activities (notebook, JAR, Python) in Azure Data Factory if the target cluster in Azure Databricks isn't running when the cluster is called by Data Factory
2. The Databricks activity will fail in Azure Data Factory – you must always have the cluster running
3. Simply add a Databricks cluster start activity before the notebook, JAR, or Python Databricks activity
4. If the target cluster is stopped, Databricks will start the cluster before attempting to execute
5. Cluster will be dynamically invoked in response to user request

Answer: b

1. What allows Azure to terminate the databricks cluster after a specified number of minutes of inactivity
2. Auto-stop
3. Auto-scales
4. Auto-termination
5. Spark Job Completion

Answer: c

1. What is “Data Plane” in Databricks cluster
2. HDFS+YARN
3. Hive Metastore+DBFS
4. Azure BLOB+Runtime Engine
5. SQL libraries for ACID Compliance

Answer: b