Q1.Which of the following is the foundation of MapReduce operations?

A. Powerful processors
B. Big Data
C. Key/value data
D. All of the above

Correct Option:C
EXPLANATION : Using key/value data as the foundation of MapReduce operations allows a powerful programming model that can be applied by a framework like Hadoop.

Q2.Which of the following commands can be used to run a WordCount program on a local Hadoop cluster?

A. $ hadoop jar wc1.jar WordCount ABC.txt output
B. $ hadoop wc1.jar WordCount ABC.txt output
C. $ hadoop jar wc1.jar WordCount ABC.txt
D. None of the above

Correct Option:A
EXPLANATION : $ hadoop jar wc1.jar WordCount ABC.txt can be used to run the program.

Q3.Which command is used to check the status of all daemons running in HDFS?
SELECT THE CORRECT ANSWER

A. fsck
B. distcp
C. jps
D. None of the above

Correct Option:C
EXPLANATION : JPS command is used to check all the Hadoop daemons like NameNode, DataNode, ResourceManager, and NodeManager running on the machine.

Q4.What will be the ouptut of "val c" in the code given below: scala> val a = Array(1,2,3,11,4,12,4,5) scala> val b = Array(6,7,4,5) scala> val c = a.toSet diff b.toSet

A. (1,11,12,3)
B. Set(1,2,11,3)
C. Set(1,2,12,3,11)
D. Set(1,2,12,4,7)

Correct Option:C
EXPLANATION ; val c will be Set(1,2,12,3,11).

Q5._____ is used to list the blocks that make up each file within the filesystem.

A. hdfs fchk / -files -blocks
B. hdfs fsck / -blocks -files
C. hdfs fsck / -files -blocks
D. hdfs fchk / -blocks -files

Correct Option : C
EXPLANATION : The fsck command is used by the administrator to check the file system on Hadoop.

Q6.The number of map tasks for a given job is driven by _____.

A. mapred.map.tasks parameter
B. Size of the data
C. Number of input splits
D. None of the above

Correct Option:C
EXPLANATION : For each input split, a map task is spawned. So, over the lifetime of a MapReduce job the number of map tasks is equal to the number of input splits.

Q7.Where is table data stored in Apache Hive by default?

A. $HIVE_HOME/config
B. $HIVE_HOME/default/user/conf/hive/warehouse
C. hdfs: //namenode_server/user/hive/warehouse
D. None of the above

Correct Option:C
EXPLANATION:All table data is stored under the default directory /user/hive/warehouse/<database name>/<table>/filename.

Q8.Which of the following command can be used to remove the error "FAILED ERROR IN SEMANTIC ANALYSIS: Dynamic partition strict mode requires at least one static partition column"?

A. SET hive.exec.dynamic.partition = true
B. SET hive.exec.dynamic.partition.mode = nonstrict
C. A and B
D. None of the above

Correct Option:C
EXPLANATION : To remove the given error one has to execute both the commands.

Q9.Read the code given below: s = 'Hi hi hi bye bye bye word count' sc.parallelize(seq).map(lambda word: (word, 1)).reduceByKey(add).collect() Output: [('count', 1), ('word', 1), ('bye', 3), ('hi', 2), ('Hi', 1)] Select the correct which will produce the above output.

A. seq = s.split()
B. seq = s.split([])
C. seq=split().s
D. None of the above

Correct Option:A
EXPLANATION : seq = s.split() will produce the given output.

Q10.The command to import only USA cities from the table "cities" is given by:

A. sqoop import --connect jdbc:mysql://mysql.example.com/sqoop --username sqoop --password sqoop --table cities -- "country = 'USA'"
B. sqoop import --connect jdbc:mysql://mysql.example.com/sqoop --username sqoop --password sqoop --table cities --where "country = 'USA'"
C. sqoop import --connect jdbc:mysql://mysql.example.com/sqoop --username sqoop --password sqoop --table city -- "country = 'USA'"

D. sqoop import --connect jdbc:mysql://mysql.example.com/sqoop --username sqoop --password sqoop --table city -- where "country" = 'USA'

Correct Option:B
EXPLANATION : We use the command-line parameter --where to specify a SQL condition.

Q11.How can we batch multiple-insert statements together in Sqoop?

A. Enable JDBC batching using the --batch parameter
B. Use "sqoop.export.records.per.statement" to specify the number of records that will be used in each insert statement
C. Use "sqoop.export.statements.per.transaction" to set how many rows will be inserted per transaction
D. All of the above

Correct Option:D
EXPLANATION : Tailored for various databases and use cases, Sqoop offers all the options for inserting more than one row at a time.

Q12.Which data extraction tool can be used to extract streaming data from social media sites in Hadoop?

A. Sqoop
B. Flume
C. Hive
D. Impala

Correct Option:B
EXPLANATION : Apache Flume is the preferred tool for efficiently collecting, aggregating, and moving large amounts of log data.

Q13.After importing MySQL data into Hive with Sqoop, you get few rows which contain escape characters "\t" and "\n". One needs to drop or replace these characters to get Hive-compatible text data. Select the command which can be used to achieve this.

A. sqoop import --connect jdbc:mysqll//localhost:3306/mysqldb --username user --password pwd --table mysqltbl --hive-import --hive-overwrite --hive-table hivedb.hivetbl -m 1 --hive-drop-import-delims --null-string '\\N' --null-non-string '\\N'
B. sqoop import --connect jdbc:mysqll//localhost:3306/mysqldb --username user --password pwd --table mysqltbl --hive-import --hive-overwrite --hive-table hivedb.hivetbl -m 1 --hive-delims-replacement " " --null-string '\\N' --null-non-string '\\N'
C. Both A and B
D. None of the above

Correct Option:C
EXPLANATION:You can use the "--hive-drop-import-delims" option to drop characters on import to give Hive-compatible text data or use "--hive-delims-replacement" option to replace characters with a user-defined string on import to give Hive-compatible text data.

Q14.Which of the following commands can be used to filter a spark dataframe, where the date is greater than "2019-02-14" and the date column is of StringType?

A. data.filter(data("date").lt(lit("2019-02-14")))
B. data.filter(data("date").gt(lit("2019-02-14")))
C. data.filter(to_date(data("date")).gt(lit("2019-02-14")))
D. data.filter(to_date(data("date")).lt(lit("2019-02-14")))

Correct Option : C
EXPLANATION : data.filter(to_date(data("date")).gt(lit("2019-02-14"))) will produce the correct output.

Q15.Given a "employee2" table: first_name string last_name string Program from pyspark.sql import HiveContext sqlContext = HiveContext(sc) < missing code> employee2.collect(); Select the missing Spark SQL query in Python which reads "Employee2" table and prints all the rows and individual column values.

A. employee2 = sqlContext.sql(select * from employee2)
B. employee2 = sqlContext(select * from employee2)
C. employee2 = sqlContext.sql("select * from employee2")
D. INSERT INTO ctas_test.employee2 SELECT * FROM employee

Correct Option : C
EXPLANATION : employee2 = sqlContext.sql("select * from employee2") will produce the correct output.

Q16.Which command is used to check the status of all daemons running in HDFS?

A. None of the above
B. jps
C. distcp
D. fsck

Correct Option : B
EXPLANATION : JPS command is used to check all the Hadoop daemons like NameNode, DataNode, ResourceManager, and NodeManager running on the machine.

Q17.Read the below code snippet: lines = sc.parallelize(['Let us do something interesting,','but do not speak about it.']) M = lines.map(lambda x: x.replace(',7 ').replace('.',' 'J.replaceC-V ').lower()) r2 = r1.flatMap(lambda x: x.split()) r3 = r2.map(lambda x: (x, 1)) r5 = r4.map(lambda x:(x[1],x[0])) r6 = r5.sortByKey(ascending=False) r6.take(20) Select the correct code snippet for which will produce the desired output shown below. [(2, 'us'), (2, 'do'), (2, 'something'), (1, let'), (1, 'about1), (1, 'it1), (1, 'do'), (1, 'but')]

A. r4 = r3.reduceByKey(lambda x,y:2x+y)
B. r4 = r3.reduceByKey(lambda x,y:x-y)
C. r4 = r3.reduceByKey(lambda x,y:x+y)
D. r4 = r3.reduceByKey(lambda x,y:x*y)

Correct Option:C
EXPLANATION : r4 = r3.reduceByKey(lambda x,y:x+y) will produce the correct output.

Q18.Read the below code snippet: val a = sc.parallelize(List("dog", "tiger", "lion", "cat", "panther", "eagle"), 2) val b = a.map(x => (x.length, x)) < Select the correct code snippet for which will produce the desired output shown below. Array[(Int, String)] = Array((4,lion), (3,dogcat), (7,panther), (5,tigereagle))

A. b.foldByKey("")(_ - _).collect
B. b.fold("")(_ + _).collect
C. b.fold("")(_ - _).collect
D. b.foldByKey("")(_ + _).collect

Correct Option:D
EXPLANATION : b.foldByKey("")(_ + _).collect will produce the correct output.

Q19.Read the below scenario: The Big data team at Northwest Airlines has built a machine learning model which predicts whether an airline flight will arrive on-time or late. The team wants to deploy this model, and also train and update it, in real-time based on the incoming data. This will continually improve predictions based on the updated data. Which of the following tools is best suited for the above scenario?

A. Spark MLlib
B. Flume
C. Kafka Steams API
D. Impala

Correct Option:C
EXPLANATION : Apache Kafka is best suited for this task.

Q20.What will be the output of the following command: sc.parallelize([3,4,5]).flatMap(lambda x: [x, x*x]).collect()

A. [[3, 9], [4, 16], [5, 25]]
B. [1, 2, 1, 2, 3, 1, 2, 3, 4]
C. [3, 9, 4, 16, 5, 25]
D. [3, 4, 16, 5, 25]

Correct Option:C
EXPLANATION : [3, 9, 4, 16, 5, 25] is the correct output.

Q21.Which of the following is not an execution mode in Pig?

A. Tez mode
B. Mapreduce mode
C. Spark Mode
D. None of the above

Correct Option:D
EXPLANATION:You can run execute Pig statements and commands using any of the modes.

Q22. Read the below code snippet: val a = sc.parallelize(List("dog", "tiger", "lion", "cat", "spider", "eagle"), 2) val b = a.keyBy(_.length) val c = sc.parallelize(List("ant", "falcon", "squid"), 2) val d = c.keyBy(.length) Select the correct code snippet for which will produce the desired output shown below. Array[(lnt, String)] = Array((4,lion))

A. c.subtractByKey(c).collect subtractByKey [Pair]
B. b.subtractByKey(c).collect subtractByKey [Pair]
C. c.subtractByKey(b).collect subtractByKey [Pair]
D. b.subtractByKey(d).collect subtractByKey [Pair]

Correct Option:D
EXPLANATION: b.subtractByKey(d).collect subtractByKey [Pair] will produce the correct output.

Q23.You have to process 20,000 200KB files in Hadoop. Select the appropriate technique that will process these files in the shortest time.

A. Mapreduce
B. Use a Sequence file
C. Both A and B
D. None of the above

Correct Option:B
EXPLANATION:Map tasks process one block of input at a time. More the number of files, the more number of map tasks are needed, resulting in greater job run-times. SequenceFile puts each small file to a larger single file and is also suitable for MapReduce.

Q24.Hive specific commands can be run from Beeline, when the Hive _____ driver is used.

A. ODBC-JDBC
B. ODBC
C. JDBC
D. All of the above

Correct Option:C
EXPLANATION:Hive specific commands can be run from Beeline, when the Hive JDBC driver is used.

Q25.values = sc.parallelize(range(10), 3) Select the correct code snippet for which will produce the desired output shown below. Output: [[0, 1, 2], [3, 4, 5], [6, 7, 8, 9]]
SELECT THE CORRECT ANSWER

A. values.collect()
B. values.glom().collect()
C. values.glom(3).collect()
D. values.collect(3)

Correct Option:B
EXPLANATION:values.glom().collect() will produce the correct output.

Q26._____ jobs are optimized for scalability but not latency.

A. Oozie
B. MapReduce
C. Hive
D. Drill

Correct Option:C
EXPLANATION:Hive jobs are optimized for scalability but not latency.

Q27.What does the following code do? lines = sc.textFile("data.txt") pairs = lines.map(lambda s: (s, 1)) counts = pairs.reduceByKey(lambda a, b: a + b)

A. The reduceByKey sorts the key-value pairs alphabetically.
B. The reduceByKey operation on key-value pairs counts how many times each line of text occurs in the file.
C. Both A and B
D. None of the above

Correct Option:B
EXPLANATION : The given code counts how many times each line of text occurs in the file.

Q28.You are required to create an RDD containing only lines that are requests for pdf files. Select the correct command which will accomplish this.

A. var pdflogs = logs.filter(line => line.contains("pdf"))
B. var pdflogs = logs.filter(line => line.contains(".pdf"))
C. var pdflogs = logs.filter(line => line.contains(".pdf"))
D. var pdflogs = logs.filter(line => line.contains(.pdf))

Correct Option:B
EXPLANATION:var pdflogs = logs.filter(line => line.contains(".pdf")) can be used to accomplish this.

Q29.Which among the following is the most popular NoSQL database for scalable big data store with Hadoop ?

A. MongoDB
B. HBase
C. Casssandra
D. None of the above

Correct Option:B
EXPLANATION : HBase is a distributed, scalable big data store that lets you host very large tables.

Q30.Read the given below code: val a = sc.parallelize(List("dog", "tiger", "lion", "cat", "panther", "eagle"), 2) b.values.collect Select the correct code snippet for which will produce the desired output shown below. Array[String] = Array(dog, tiger, lion, cat, panther, eagle)

A. val b = a.map(x => (x.length, x*y))
B. val b = a.map(x => (x.length, x+y))
C. val b = a.map(x => (x.length, x))
D. val b = a.map(x => (x.length, x-y))

Correct Option:C
EXPLANATION:val b = a.map(x => (x.length, x)) will produce the given output.