1.which of the following statements about the spark driver is true?

Ans: The Spark driver is responsible for scheduling the execution of data by various worker nodes in cluster mode.

2. which of the following describes the relationship between nodes and executors?

Ans: An executor is a processing engine running on a node.

3. which of the following statements about the slots is incorrect?

Ans: Slots are the most granular level of execution in the spark execution hierarchy.

4. which of the following describes a task?

Ans: A task is a combination of a block of data and set of transformers that will run on a single executor.

5.which of the following statements about the spark jobs is incorrect?

Ans: Jobs are collections f tasks that are divided based on hen language variables are defined.

6. which of the following operations is most likely to result in a shuffle?

Ans: **DataFrame.join()**

7.DataFrame df is very large with a small number of partitions, fewer than there are executors in the cluster. Based on this situation, which of the following is incorrect? Assume there is one core per executor.

Ans: Performance will be suboptimal because not all executors will be utilized at the same time.

8. which of the following is the most complete description of lazy evolution?

Ans: A process is lazily evaluated if its execution does not start until it is put into action by some type of trigger

9. which of the following DataFrame operations is classified as action?

Ans: **DataFrame.take()**

10. which of the following DataFrame operations is classified as a wide transformation?

Ans: **DataFrame.join()**

11. Which of the following describes the difference between cluster and client execution modes?

Ans: The cluster execution mode runs the driver on worker node within a cluster, while the client execution mode runs the driver on the client machine(also known as a gateway machine or edge node).

12. which of the following cluster configurations will fail to ensure completion of a spark application in light of worker node failure?

Note: each configuration has roughly the same compute power using 100GB of Ram and 200 cores

Ans: They should all ensure completion because worker nodes are fault tolerant.

13. which of the following cluster configurations is least likely to experience an out-of-memory error in response to data skew in a single partition?

Note: each configuration has roughly the same compute power using 100GB of Ram and 200 cores

Ans: Scenario #6

14. Which of the following best describe the similarities and differences between the MEMORY_ONLY storage level and the MEMORY_AND_DISK storage level?

Ans: The MEMORY_ONLY storage level will store as much data as possible in memory and will recompute any data that does not fit in memory as it's called.

The MEMORY_AND_DISK storage level will store as much data as possible in memory and will store any data that does on fit in memory on disk and read it as it's called.

15. Which of the following spark properties is used to configure the maximum size of an automatically broadcasted DataFrame when performing a join?

Ans: **spark.sql.autoBroadcastJoinThreshold**

16. Which of the following operations can be used to create a new DataFrame that has 12 partitions from an original DataFrame df that has 8 partitions?

Ans: **df.repartition(12)**

17. which of the following allows for parallel execution to be performed on spark DataFrames?

Ans: RDDs

18. which of the following operations can be used to create DataFrame with a subset of columns from DataFrame storesDF that are specified by name?

Ans: **storesDF.select()**

19. The code block shown below should return a DataFrame containing all columns from DataFrame storesDF except for column sqft and column customerSatisfaction. choose the response that correctly fills in the numbered bank within the code block to complete this task?

Ans: **storesDF.drop("sqft","customerSatisfaction")**

20. which of the following code blocks returns a DataFrame containing only the rows from DataFrame storesDF where the value in column sqft is less than or equal to 25,000?

Ans: **storesDF.filter(col("sqft") <= 25000)**

21. which of the following code blocks returns a DataFrame containing only the rows from DataFrame storesDF where the value in column sqft is less than or equal to 25,000 AND the value in column customerSatisfaction is greater than or equal to 30?

Ans: **storesDF.filter((col("sqft") <= 25000) & (col("customerSatisfaction") >=30))**

22. The code block shown below should return a new DataFrame from DataFrame storesDF where column storeID is of the type string. choose the response that correctly fills in the numbered blanks within the code block to complete this task?

Ans: **storesDF.withColumn("storeID",col("storeID").cast(StringType())**

23. which of the following code blocks returns a new DataFrame with a new column customerSatisfactionAbs that is the absolute value of column

customerSatisfaction in DataFrame storesDF? Note that column customerSatisfactionAbs is not in the original DataFrame storesDF.

Ans:
storesDF.withColumn("customerSatisfactionAbs",abs(col("customerSatisfaction")))

24. The code block shown below should return in new DataFrame from DataFrame storesDF where column numberOfManagers is the constant integer 1. choose the response that correctly fills in the numbered blanks within the code block to complete this task?

Ans: **storesDF.withColumn("numberOfManagers",lit("1"))**

25. The code block shown below contains an error. the code block is intended to return a new DataFrame where column managerName from DataFrame storeDF is split at the space character into column managerFirstName and managerLast Name. Identified the error.

A sample of DataFrame storeDF is displayed below

```
(storesDF.withColumn("managerFirstName",col("managerName").split(" ")[0])
        .withColumn("managerLastName",col("managerName").split(" ")[1]))
```

Ans: The split() operation comes from the imported functions object. It accepts a column object and split character as arguments it is not a method of a column object

26. Which of the following code blocks return a new DataFrame where column productCategories only has one word per row, resulting in a DataFrame with many more than DataFrame storesDF?

A sample of storesDF is displayed below

Ans:
storesDF.withColumn("productCategories",explode(col("productCategories")))

27. which of the following code block returns a new DataFrame with column storeDescription where the pattern "Description" has been removed from the beginning of column storeDescription in DataFrame storeDF?

Ans:
storesDF.withColumn("storeDescription",regexp_replace("storeDescription", "^Description:',""))

28. The code block shown below should return a new DataFrame where column division from DataFrame storesDF has been renamed to column state and column managerName from DataFrame storesDF has been renamed to column managerFullName. choose the response that correctly fills in a numbered blank within the code block to complete this task.

Ans:
storesDF.withColumnRenamed("division","state").withColumnRenamed("managerName","managerFullName")

29. The code block shown below should return a new DataFrame where rows in DataFrame storesDF containing at least one missing value have been dropped. choose the response that correctly fills in the numbered blanks within the code block to complete this task?

Ans: **storesDF.na.drop(how="any")**

30. which of the following code block returns DataFrame where every Row is unique?

Ans: **storesDF.distinct()**

31. which of the following code blocks will most quickly return in approximation for the number of distinct values in column division in DataFrame storesDF?

Ans:
storesDF.agg(approx_count_distinct(col("division",0.01).alias("divisionDistinct"))

32. which of the following operations calculate the simple average of group of values, like a column?

Ans: **mean()**

33. which of the following operations can be used to return the number of rows in a DataFrame?

Ans: **DataFrame.count()**

34. Which of the following code blocks fails to return the number of rows in DataFrame storesDF for each distinct combination of values in column division and column storeCategories?

Ans: **storesDF.groupBy("division").groupBy("storeCategory).count()**

35. The code block shown below should return a collection of summary statistics for column sqft in DataFrame storesDF. choose the response that correctly fills in the numbered blanks within the code block to complete this task?

Ans: **storesDF.describe("sqft")**

36. which of the following code blocks fails to return a DataFrame sorted alphabetically based on column division?

Ans: **storesDF.sort(asc("division"))**

37. which of the following code blocks returns a 10% sample of rows from DataFrame storesDF with replacement?

Ans: **storesDF.sample(True, fraction=0.1)**

38. Which of the following code block returns all the rows from DataFrame storesDF?

Ans: **storesDF.collect()**

39. which of the following code blocks applies the function accessesPerformance() to each row of DataFrame storesDF?

Ans: **[assessPerformance(row) for row in storesDF.collect()]**

40. the code block shown below should print the schema all DataFrame storesDF. Choose the response correctly fills in the numbered blanks within the code block to complete this task?

Ans: **storesDF.printSchema(Nothing)**

41. which of the following code block creates and registers a SQL UDF named "ASSESS_PERFORMANCE" using the Python function accesPerformance() and applies it to column customerSatisfaction table to stores?

Ans: **spark.udf.register("ASSES_PERFORMANCE", assessPerformance)**

spark.sql("SELECT customerSatisfaction, ASSES_ PERFORMANCE(customerSatisfaction) AS result FROM stores")

42. The code block shown below contains an error. the code block is intended to create a Python UDF assesPerformanceUDF() using the integer returning Python function assessPerformance() and apply to column customerSatisfaction in DataFrame storesDF identify the error

Code block:

- assesPerformanceUDF = udf(assessPerformance)
  storesDF.withcolumn("result",assesPerformanceUDF(col("customerSatisfaction"))

Ans: **The return type of the assesPerformanceUDF() is not specified in the udf() operation**.

43. which of the following code blocks users SQL to return a new DataFrame containing column storeID and column managerName from a table created from DataFrame storesDF?

- Ans: **storesDF.createOrReplaceTempView("stores")**
  Spark.sql("SELECT storeID, managerName FROM stores")

44. The code block shown below should create a single column DataFrame from Python list years which is made-up of integers. choose the response that correctly fills in the numbered blanks within the code block to complete this task?

Ans: **spark.createDataFrame(years,IntegerType())**

45. which of the following code block attempts to cache the partitions of DataFrame storesDF only in spark's memory?

Ans: **storesDF.persist().count()**

46. which of the following code block returns will always return a new 4 partition DataFrame from the 8 partition DataFrame storesDF without inducing a shuffle?

Ans: **storesDF.coalesce(4)**

47. The code block shown below should return a new 12 partition DataFrame from DataFrame storesDF. choose the response that correctly fills in the numbered blanks within the code block to complete this task?

Ans: **storesDF.repartition(12)**

48. which of the following spark properties is used to configure whether DataFrame partitions that do not meet a minimum size threshold are automatically coalesced into larger partitions during a shuffle?

Ans: **spark.sql.adapative.coalescePartitions.enabled**

49. which of the following code block returns a DataFrame containing a column openDateString a string representation of javas SimpleDateFormat?

Note that column openDate is of type integer and represents a date in the Unix epoch format the number of seconds since midnight on January 1st,1970.

An example of Javas SimpleDateFormat is "Sunday, Dec 4, 2008 1:05PM"

Ans:

storesDF.withColumn("openDateString",from_unixtime(col("openDate"), "EEEE, MMM d, yyyy h:mm a"))

50. The code block shown below contains an error. the code block intended to return a DataFrame containing a column dayOfyear, an integer representation of the day of the year from column openDate from DataFrame storesDF. identify the error

Note that column openDate is of type integer and represents a date in the Unix epoch format the number of seconds since midnight on January 1st,1970.

A sample of StoresDF is Displayed below:

Code block:

```
storesDF.withColumn("dayOfyear", dayOfyear(col("openDate"))
```

Ans: **The dayOfyear() operation cannot extract the day of year from a column of type integer column openDate must first be converted to type Timestamp**

51. The code block shown below should return a new DataFrame that is the result of an inner join between DataFrame storeDF and DataFrame employeesDF on column storeID. choose the response that correctly fills in the numbered blank within the code block to complete this task.

Ans: **storesDF.join(employeesDF,"storeId","inner")**

52. The code block should return a new DataFrame that the result of an outer join between DataFrame storeDF and DataFrame employeeDf on column storeId. choose the response that correctly fills in the number blank within a code block to complete this task?

Ans: **storesDF.join(employeeDf,"storeId","outer")**

53. which of the following pairs of arguments cannot be used in DataFrame.join to perform an inner join on two DataFrames. named and aliased with "a" and "b" respectively. to specify two key columns?

Ans: **on=["column1", "column2"]**

54. The code block shown below should efficiently perform a broadcast join of DataFrame storesDF and much larger DataFrame employeesDF using key column storeID. choose the response that correctly fills in the numbered blank within the code block to return complete this task.

Ans: **employeesDF.join(broadcast(storesDF), "storeId")**

55. The code Block shown below contains an error. the code block is intended to return a new DataFrame that is the result of cross join between DataFrame storeDF and DataFrame employeeDf. identify the errors.

Code block:

storesDF.join(employeesDF,"cross")

Ans: A cross join is not implemented by the DataFrame.join() operation—the DataFrame.crossJoin() Operation should be used instead.

56. the code block shown below contains an error. the code block is intended to return a new DataFrame that is the result of a position wise union between DataFrame storeDF and DataFrame acquiredstoresDF. identify the error

Code block:

storesDF.unionByName(acquiredStoresDF)

Ans: The UnionByName() Operation is a standalone operation rather than a method of DataFrame—It should have both DataFrames as arguments.

57. which of the following code block writes a DataFrame storeDf to file path filepath as CSV?

Ans: **StoresDf.write.csv(filePath)**

58. The code block shown below should write DataFrames storesDF is file path filepath as parquet and partition by values in column division. Choose the response correctly fills in the numbered blanks within the code block to complete this task?

Ans: **storesDF.write.partitionBy("division").parquet(filepath)**

59. the code block show below should read a parquet at the file path filepath into a DataFrame. choose the response that correctly fills in the numbered blanks within the code back to complete this task?

Ans: **spark.read.load(filepath,source="parquet")**

60. the code block shown below should read a JSON at the file path filepath into a DataFrame with the specified schema schema. choose the response that correctly fills in the number blacks within the code plug to complete this task?

Ans: **spark.read.schema(schema).format("json").load(filePath)**

1.**Which of the following describes the Spark driver?**
Ans: The Spark driver is responsible for performing all executing in all execution modes-- It is entire spark application.

2.Which of the following describes the relationship between nodes and executors?
Ans: There are always the same number of executors and nodes.

3.Which of the following is the most granular level of the spark execution hierarchy?
Ans: Task is the Granular level

4.Which of the following types of processes induces a stage boundary?

Ans: Shuffle induces a stage boundary.

5.Which of the following cluster configurations will induce the least network traffic during a shuffle operation?

Answer: It depends on executors and memory partition One Driver with one Executor it wont work. one driver with 2 executor will have least traffic. so ans of this question is one driver with 2 executor.

6.Which of the following is the most complete description of lazy evaluation?

Ans: A process is Lazily evaluated if its execution does not start until it is place into action by some type of trigger

7.Which of the following DataFrame operations is classified as an action?

Answer: DataFrame.take()

8.Which of the following DataFrame operations is classified as a wide transformation?

Answer: DataFrame.join()

9.Sparks execution/deployment mode determines where the driver and executors are physically located when a spark application is run. Which of the following

spark execution/deployment modes does not exist? If they all exist, please indicate so with Response E.

Answer: Standard mode

10: Which of the following statements about spark's stability is incorrect?

Answer: Spark will reassign the driver to a worker node if the driver's node fails.

11.Which of the following cluster configurations is most likely to experience delays due to garbage collection of a large DataFrame?

Answer: Answer of this question depends on How much the executors and cores are distributed. More number of executors and more number of cores will phase that type of experience.

1 drive 8 executors 12.5 GB and 25 cores per Executor is the best ans.

12.Which of the following describes slots?

Answer: Slots are resoure threads that can be used for parallelization within a spark application.

14.Which of the following situations, in which will it be most advantageous to store DataFrame df at the MEMORY_AND_DISK storage level rather than the MEMORY_ONLY Storage level?

Answer: When its faster to read all the computed data in DataFrame df that cannot fit into memory from disk rather than recompute it based on its logical.

15.A Spark Application has a 128 GB DataFrame A and a 1 GB DataFrame B. If a broadcast join were to be performed on these two DataFrames, Which of the following describes which DataFrame should be broadCasted and why?

Ans: DataFrame B should be broadcasted because it is smaller and will eliminate the need for the shuffling of itself.

20.Which of the following operations can be used to create a new DataFrame that has 12 partitions from an original DataFrame df that has 8 partitions?

Answer: df.repartition(12)

21.Which of the object types cannot be contained within a column of a spark DataFrame?

Ans: Null

22.Which of the following operations can be used to create a DataFrame with a subset of columns from DataFrame storesDF that are specified by name?

Answer: storesDF.select()

23.Which of the following describes partitions?

Answer: A Partition is an automatically-sized segment of data that is used to create efficient logical plans.

24.The code block shown below should return a DataFrame containing all columns from DataFrame storesDF except for column "sqft" and column "customerSatiFaction". Choose the response that correctly fills in the numbered blanks within the code block to complete this task.

Answer: storesDF.drop("sqft", "customerSatisFaction")

25.which of the following code blocks returns a DataFrame containing on the rows from DataFrame storesDF where the value in column sqft is less than or equal to 25,000?

Answer: storesDF.filter(col("sqft") >= 25000)

26. Which of the following code blocks returns a DataFrame containing only the rows from DataFrame storesDF where the value in column sqft is less than or equal to 25,000 AND the value in column customerSatisfaction is greater than or equal to 30?

Ans: storesDF.filter((col("sqft") >= 25000) & (col("customerSatisfaction") >= 30))

27.The code block shown below should return a new DataFrame where column division from DataFrame storesDF has been renamed to column state and column managerName from DataFrame storesDF has been renamed column

managerFullName. Choose the response that correctly fills in the numbered blanks within the code block to complete this task.

Ans:

```
storesDF.withColumnRenamed("division","state").withColumnRenamed("manage
rName", "managerFullName")
```

28.Which of the following code blocks returns a new DataFrame with a new column "employeesPerSqft" that is the quotient of column "numberOfEmployees" and column "sqft", both of which are from DataFrame "storesDF"? Note that column employeesPerSqft is not in the original DataFrame storesDF.

Answer:

```
storesDF.withColumn("employeesPerSqft",col("numberOfEmployees")/col(
"sqft"))
```

29.Which of the following code blocks returns a new DataFrame from DataFrame storesDF where columnn modality is the constant string "PHYSICAL"? Assume DataFrame storesDF is the only defined language variable.

Answer: storesDF.withColumn("modality", lit("PHYSICAL"))

30.Which of the following code blocks returns a DataFrame where column managerName from DataFrame storesDF is split at the space character into column managerFirestName and column managerLastName?

Answer: storeDF.withColumn("managerFirstName", split("managerName", " ")[0])

.withColumn("managerLastName", split("managerName", " ")[1])

31.The code block shown below should return a new DataFrame where column productCategories only has one word per row, resulting in a DataFrame with many more rows than DataFrame storesDF. Choose the response that correctly fills in the numbered blanks within the code block to complete this task.

Answer: storesDF.withColumn("prodctCategories",
explode(col("productCategories"))

32. Which of the following code blocks returns a new DataFrame with column storeDescription where the pattern "Description:" has be removed from the beginning of column storeDescription in DataFrame storesDF?
Answer: storesDF.withColumn("storeDescription", regex_replace("storeDescription", "^Description: ", ""))

33.The code block shown below should return a new DataFrane where rows in DataFrane storesDF containing at least one missing value have been dropped. Choose the response that correctly fills in the numbered blanks within the code block to complete this task.

Answer: storesDF.na.drop(how = "any")

34. Which of the following code blocks returns a DataFrane where every row is unique?

Answer: storesDF.distinct()

35. The code block shown below contains an error. The code block is intended to return the exact number of distinct values in division in column division in DataFrame storesDF. Identify the error.

Answer: The apporx_count_distinct() operation needs a second argument to set the "rsd" parameter to ensure it returns the exact number of distinct values.

36. Which of the following operations calculates the simple average of a group of values, like a column?

Answer: mean()

37. Which of the following operations can be used to return the number of rows in a DataFrame?

Answer: DataFrame.count()

38. Which of the following operations returns a groupedData object?

Answer: DataFrame.groupBy()

38. Which of the following code blocks returns a collection of summary statistics for all columns in 'DataFrame storesDF?

Answer: storesDF.describe()

39. Which of the following code blocks returns a DataFrame sorted in ascending order (with missing values fist) based on column sqft?

Answer: storesDF.orderBy(asc_nulls_first(col("sqft")))

40.The code block shown below should return a 25 percent sample of rows from DataFrame storesDF with reproducible results. Choose the response that correctly fills in the numbered blanks within the code block to complete this task.

Answer: storesDF.sample(fraction = 0.25,seed = True)

41.Which of the following operations can be used to return all of the rows from DataFrame?

Answer: DataFrame.collect()

42. The code block shown below should apply the function assessPerformance() to each row of DataFrame storesDF. Choose response that correctly fills in the numbered blanks within the correct block to complete this task.

Answer:

43. The code block shown below contains an error. The code block is intended to print the scema of DataFrame storesDF. Identify the error.

Answer: The printSchema member of DataFrame is an operation and need to be follwed by parentheses.

44. The code block shown below should create and register a SQL UDF named "ASSESS_PERFORMANCE" using the python function assessperformance() and

apply it to column customerSatistfaction in table stores. Choose the response that correctly fills in the numbered blanks within the code block to complete this task.

```
spark.__1__.__2__(__3__,__4__)
spark.sql("SELECT customerSatisfaction, __5__(customerSatisfaction) As result FROM stores")
```

Answer : spark.udf.register("ASSESS_PERFORMANCE", assesPerformace)
spark.sql("SELECT customerSatisfaction, ASSESS_PERFORMANCE(customerSatisfaction)")

45. The code block shown below contains an error. The code block is intended to create a python UDF assessPerformaceUDF() using the integer-returning python function assessPerformance() and apply it to column customerSatisfaction in DataFrame storesDF. Identiy the error.

```
assessPerformanceUDF = udf(assesPerformance)
storesDF.withColumn("result", assessPerformanceUDF(col("customerSatisfaction")))
```

Answer : The assessPerformance() operation is not properlly registerd as a UDF.

46. The code block shown below contains an error. The code block is intended to use SQL to return a new DataFrame containing column sotredId and column managerName from a table created from DataFrame storesDF. identify the error.

Answer: The createOrReplaceTempView() operation should be accessed via the spark variable rather than DataFrame storesDF.

47. The code block shwn below contains an error. The code block is i ntended to create a single-column DataFrame from python loist year which is made up of integers. Identify the error.

```
code block: spark.createDataframe(years, IntegerType)
```

Answer: The years list should be wrapped in another like [years] to make clear that it is a column rather than a row.

48. Which of the following code blocks attempts to cache the partitions of DataFrame storesDF only in sparks memory?

Answer: storesDF.persist().count()

49. which of the follwing operations can be used to return a new DataFrame from DataFrame storesDF without inducing a shuffle?

Answer: storesDF.coalesce(1)

50. The code block shown below contains an error. The code block is intended to return a new 12 Partition DataFrame from the 8 Partition DataFrame storesDF by inducing a shuffle. Identify the error.

code block: storesDF.coalesce(12)

Answer: The coalesce() operation does not induce a shuffle and cannot increase the number of partitions -- The repartition() operation should be used instead.

51.The code block shown below should adjust the number of partitions used in wide transformations like join() to 32. Choose the response that correctly fills in the numbered blanks within the code block to complete this task.

1(2,3)

Answer: