WEEK 12 QUIZ SOLUTION

1. Which of these notations are used to refer a column in a Dataframe/Dataset, using column object notation but is specific to scala only and not python.
    a. column("column_name"), $"column_name"
    b. col("column_name"), 'column_name
    c. *$"column_name", 'column_name
    d. column("column_name"),col("column_name")

    Explanation: *$"column_name", 'column_name is not supported in scala.

2. Which saveMode will give error if output folder already exist
    a. Append
    b. *ErrorIfExists
    c. Overwrite
    d. Ignore

    Explanation: **ErrorIfExists** will throw error if output folder already exists.

3. Partition pruning is possible using _____
    a. rePartition
    b. ***partitionBy**

    Explanation: partition pruning is possible with partitionBy. It is important to detect and avoid scanning data that is irrelevant to the executed query, an optimization which is known as partition pruning

4. Find odd man out
    a. select
    b. where
    c. groupBy
    d. *groupByKey

    Explanation: ***groupByKey** is odd because it's a low level transformation and all others are high level transformation

5. Which of the following is not support by spark by default
    a. json
    b. parquet
    c. *avro
    d. csv

    Explanation: ***avro** is not supported by default bu you can add avro jar explicitly and work on it.

6. If two files contains same column name and you are performing join using dataframe then which of the following is true.
    a. *it will throw error column name is ambiguous
    b. After the join, rename the ambiguous column in one of the dataframe

c. Before the join, drop the column from one table.
d. It will execute and show output.

Explanation: ***it will throw error, column name is ambiguous** because it gets confused which column to display. To solve this problem there are two ways

- Before the join, rename the ambiguous column in one of the dataframe
- After the join, drop the column from one table.

7. In spark SQL provides ____ function is used to transpose row to column
   a. *pivot()
   b. groupBy
   c. unPivot()
   d. select

   Explanation: ***pivot()** function is used to transpose row to column

8. How to rename a column using dataframe in spark
   a. *withColumnRenamed
   b. renameColumn

   Explanation: ***withColumnRenamed** method is used to rename a column using dataframe

9. Which of these will create a new stage, when we create a spark application using dataframe API?
   a. show()
   b. *repartition
   c. where
   d. inferSchema()

Explanation: ***repartition** is a wide transformation hence new stage is created.

10. Convert a dataframe to a datset by _____ and dataset to dataframe by _____
    a. toDS(), Casting to Case class
    b. toDF(), toDS()
    c. *Casting to Case class, toDF()
    d. toDS(), toDF()

    Explanation: Convert a dataframe to a datset by **Casting to Case class** and dataset to dataframe by **toDF()**

11. Simple repartitioning dataframe into number of partitions has advantage of ____, but disadvantage being _____?
    a. lesser shuffling, partitions cannot be pruned
    b. equal distribution of data among partitions, lesser parallelism

c. more parallelism, largely uneven data distribution among partitions
d. *more parallelism, full shuffling of data involved & also no partition pruning

Explanation: Simple repartitioning dataframe into number of partitions has advantage of _ **more parallelism** _, but disadvantage is , it requires _ **full shuffling of data involved & also no partition pruning** ___

12. Which metastore when used in spark, will ensure metadata remains intact even if spark application terminates.
    a. Spark's catalog metastore
    b. ***Hive Metastore**

    Explanation: ***Hive Metastore will store your metadata even if spark terminates.**

13. To add a new column to a dataframe we use:
    a. toColumn
    b. *withColumn
    c. newColumn
    d. addColumn

    Explanation: ***withColumn** is used to add new column to a dataframe**.**

14. The driver can pass the udf code directly to the executors.
    a. *TRUE
    b. FALSE

    Explanation: Driver can pass the udf code directly to the executors.

15. Point out the false statement.
    a. *dataframes are mutable
    b. withColumn can be used to update existing column in a dataframe
    c. for a dataframe column, named column1, count(column1) will exclude nulls from the result

    Explanation: ***dataframes are mutable** is false because they are immutable

16. Which of these transformations cannot be used with dataframes?
    a. groupBy
    b. join
    c. sort
    d. *reduceByKey

    Explanation: Dataframe is high level API and cant use low level API methods.

17. monotonically_increasing_id, creates values for a dataframe column which are:
    a. unique, necessarily consecutive
    b. *unique, not necessarily consecutive
    c. repeating

d. might be either unique or repetitive

Explanation: monotonically_increasing_id generates unique ids but not necessary that they are consecutive

18. Suppose we have an orders dataframe and a customers dataframe. We want to join these two so that we get all the customers who have made purchases as well as customers who have never made any purchases. Considering these below statements, what should be the joinType used here? Consider customer as left table and orders as right table.
    a. *left
    b. right
    c. inner
    d. outer

19. Suppose we have an orders dataframe and a customers dataframe, we want to find all the customers who have made purchases, and also the new customers who have made purchases, but might have not been registered yet in customers dataset. What should be the joinType used here? Consider customer as left table and orders as right table.
    a. left
    b. *right
    c. inner
    d. outer

20. Which of these is false regarding a join in dataframes? Multiple can be chosen.
    a. join is a transformation and accepts three parameters
    b. same keys should reside on same nodes for join to occur
    c. *exchange buffer resides in the driver
    d. *shuffle phase happens before exchange write

Explanation: C & D are false statement