Register Number:

DATE:

ST. JOSEPH'S COLLEGE (AUTONOMOUS), BENGALURU-27
M.Sc. BIG DATA ANALYTICS – III SEMESTER
SEMESTER EXAMINATION: OCTOBER 2021
(Examination conducted in January-March 2022)

BDA 3321: ENABLING TECHNOLOGIES FOR DATA SCIENCE

TIME: 2.5 HOURS **MAXIMUM MARKS: 70**

This Question Paper Contains FIVE Printed Papers and THREE Parts

PART A

Answer ALL questions **20 X 1 = 20**

1. Which of the following is a transformation?

A : foreach()

B : flatMap()

C : save()

D : count()

2. Which of the following is an actions

A : count()

B : printSchema()

C : cache()

D : sort()

3. Given a dataframe df, select the code that returns its number of rows:

A : df.take('all')

B : df.collect()

C : df.count()

D : df.numRows()

4. Which of the following language is not supported by Spark?

A : Java

B : Pascal

C : Scala

D : Python

5. Spark is developed in which language

A : Java

B : Scala

C : Python

D : R

6. broadcast variables are _____ and lazily replicated across all nodes in the cluster when an action is triggered

A : mutable

B : immutable

C : both

D : None of above

7. Broadcast variables are shared, immutable variables that are cached on every machine in the cluster instead of being serialized with every single task.

A : True

B : False

C : Can’t Specify

D : None

8. Spark is best suited for _____ data.

A : Real-time

B : Virtual

C : Structured

D : All of the above

9. What is action in Spark RDD?

A : The ways to send result from executors to the driver

B : Takes RDD as input and produces one or more RDD as output.

C : Creates one or many new RDDs

D : All of the above

10. Which one of the following command triggers an eager evaluation?

A : df.filter()

B : df.select()

C : df.show()

D : df.limit()

11. _____ is a distributed machine learning framework on top of Spark.

A : MLlib

B : Spark Streaming

C : GraphX

D : RDDs

12. Spark SQL provides a domain-specific language to manipulate _____ in Scala, Java, or Python.

A : Spark Streaming

B : Spark SQL

C : RDDs

D : All of the mentioned

13. Fault Tolerance in RDD is achieved using

A : Immutable nature of RDD

B : DAG (Directed Acyclic Graph)

C : Lazy-evaluation

D : None of the above

14. The shortcomings of Hadoop MapReduce was overcome by Spark RDD by

A : Lazy-evaluation

B : DAG

C : In-memory processing

D : All of the above

15. RDD is fault-tolerant and immutable

A : True

B : False

C : Both

D : None

16. Spark is engineered from the bottom-up for performance, running _____ faster than Hadoop by exploiting in memory computing and other optimizations.

A : 100x

B : 150x

C : 200x

D : None of the mentioned

17. Spark is packaged with higher level libraries, including support for _____ queries.

A : SQL

B : C

C : C++

D : None of the mentioned

18. Which of the following is true for RDD?

A : RDD is a programming paradigm

B : RDD in Apache Spark is an immutable collection of objects

C : It is a database

D : None of the above

19. Which of the following is not the feature of Spark?

A : Supports in-memory computation

B : Fault-tolerance

C : It is cost-efficient

D : Compatible with other file storage system

20. Which of the following is the reason for Spark being Speedy than MapReduce?

A : DAG execution engine and in-memory computation

B : Support for different language APIs like Scala, Java, Python and R

C : RDDs are immutable and fault-tolerant

D : None of the above

PART B

Answer ANY SIX Questions **6 X 5 = 30**

1. Explain briefly about big data characteristics

2. What are the features of pyspark

3. Explain the types of operations supported by RDDs.

4. What are the important components of the Spark ecosystem?

5. What are the different levels of persistence in Spark?

6. Explain briefly about spark architecture?

7. What is a lazy evaluation in Spark?

8. What is a Parquet file and what are its advantages?

PART

Answer Any Two Questions **2 X 10 = 20**

1. Explain lambda architecture and spark streaming architecture.

2. Explain how Spark runs applications with the help of its architecture.

3. Explain pyspark dataframe and features of pyspark sql in detail.