5/31 **0605N331** Candidate's Seat No :_____

IMBA in BI/IB/APR Sem.-8 Examination

BI

B. D. A.

Time : 2-30 Hours] **May-2025** **[Max. Marks : 70**

Instructions :(1) This paper contains **Thirty Five** questions.
(2) Each Question is of 2 Marks.
(3) Each Question is of multiple choices.
(4) All questions are compulsory.

NO.	QUESTION	Marks
Q.1	What is Hadoop primarily used for? a) Email communication b) Financial transactions c) Distributed storage and processing of Big Data d) Creating relational databases	2
Q.2	Which of the following is a transformation in Spark RDD? a) collect() b) count() c) map() d) take()	2
Q.3	Big Data Analytics helps in: a) Data deletion b) Making informed decisions c) Slowing down processes d) Decreasing data volume	2
Q.4	MongoDB is best suited for: a) Banking systems b) Real-time analytics c) Spreadsheet operations d) Desktop publishing	2
Q.5	OLAP stands for: a) Online Light Analytical Processing b) Overlapping Latency and Processing c) Online Analytical Processing d) On-premise Log Analytics Program	2

(P.T.O)

N331-2

Q.6 Which of the following is NOT a characteristic of Big Data (3Vs)? **2**

a) Volume
b) Velocity
c) Variability
d) Variety

Q.7 What is the result of applying filter(lambda x: x % 2 == 0) on [1,2,3,4,5]? **2**

a) [1,3,5]
b) [2,4]
c) [1,2,3,4,5]
d) []

Q.8 What is the primary difference between PySpark DataFrame and RDD? **2**

a) RDD is immutable, while DataFrame is mutable
b) DataFrame has a schema, while RDD does not
c) RDD supports lazy evaluation, while DataFrame does not
d) DataFrame is slower than RDD for most operation

Q.9 What does HDFS stand for? **2**

a) High Definition File Storage
b) Hadoop Distributed File System
c) Hadoop Data File Store
d) High Data File System

Q.10 What will this return? **2**

```
sc.parallelize([1, 2, 3, 4, 5]).
filter(lambda x: x % 2 == 0).
map(lambda x: x * 10).
collect()
```

a) [10, 30, 50]
b) [2, 4]
c) [20, 40]
d) [10, 20, 30, 40, 50]

Q.11 If we want only the first 2 results, we should use: **2**

a) rdd.collect()
b) rdd.first(2)
c) rdd.top(2)
d) rdd.take(2)

Q.12 Which method returns unique elements from an RDD? **2**

a) map()
b) filter()
c) distict()
d) flatMap()

Q.13 MapReduce works in how many main stages? **2**

a) 1
b) 2

c) 3
d) 4

Q.14 Which of the following is an example of an action in Spark? **2**
a) filter()
b) map()
c) collect()
d) flatMap()

Q.15 Which function gives the average salary per department? **2**
a) df.groupBy("department").sum("salary")
b) df.groupBy("salary").avg("department")
c) df.groupBy("department").avg("salary")
d) df.groupBy("department").min("salary")

Q.16 What will df.groupBy("department").count().show() display? **2**
a) Average of departments
b) Total salary by department
c) Number of rows in each department
d) Maximum age in each department

Q.17 What does the following code return? **2**
df.select("name").show()
a) All rows with all columns
b) Only the column name
c) Only the first row
d) Only rows where name is not null

Q.18 What does this return? df.agg({"salary": "avg"}).show() **2**
a) Minimum salary
b) Average salary
c) Sum of salary
d) All rows of salary

Q.19 Which Big Data technology is used for ETL operations on large datasets? **2**
a) Hive
b) Pig
c) Spark
d) Kafka

Q.20 Which of the following is an example of unstructured data? **2**
a) Excel spreadsheet
b) SQL database
c) Video File
d) Relational table

Q.21 Which method returns the schema of a DataFrame? **2**
a) df.describe()
b) df.schema()
c) df.printSchema()

(P.T.O)

N331-4

d) df.structure()

Q.22 What is the correct code to remove duplicate items? **2**

a) rdd.unique()
b) rdd.filter()
c) rdd.distinct()
d) rdd.deduplicate()

Q.23 Which code renames the column 'name' to 'employee_name'? **2**

a) df.renameColumn("name", "employee_name")
b) df.withColumnRenamed("name", "employee_name")
c) df.changeColumn("name", "employee_name")
d) df.alias("employee_name")

Q.24 What does this code do? **2**

df.filter(df.age > 30).show()

a) Filters out rows with age < 30
b) Shows all columns with null values
c) Displays only column 'age'
d) Removes duplicates

Q.25 What is the output of sc.parallelize([1,2,3]).map(lambda x: x*2).collect()? **2**

a) [1,2,3]
b) [2,4,6]
c) [3,6,9]
d) [1,4,9]

Q.26 To show the first 5 rows of the DataFrame, which command should you use? **2**

a) df.show(5)
b) df.head(5)
c) df.top(5)
d) df.take(5)

Q.27 What will rdd.count() return for this RDD: sc.parallelize([1, 2, 3, 4])? **2**

a) 3
b) 4
c) 5
d) [4]

Q.28 Which function registers a DataFrame as a temporary view for SQL queries? **2**

a) df.view("temp")
b) df.registerView()
c) df.createOrReplaceTempView("temp")
d) df.asTempTable()

N331-5

Q.29 Which join type keeps only common rows from both DataFrames? **2**

a) Left
b) Right
c) Inner
d) Outer

Q.30 In MongoDB, what is a collection? **2**

a) A single row
b) A single column
c) A group of documents
d) A group of databases

Q.31 Why would you use take(n) instead of collect()? **2**

a) To apply a transformation
b) To save all data
c) To sample a few elements
d) To shuffle the RDD

Q.32 What is the output of this code? **2**

```
sc.parallelize(["apple", "banana", "cherry"]).
map(lambda x: x.upper()).
collect()
```

a) ["apple", "banana", "cherry"]
b) ["APPLE", "BANANA", "CHERRY"]
c) ["Apple", "Banana", "Cherry"]
d) Error

Q.33 Code to get word count from RDD of lines? **2**

a) rdd.map(lambda x: (x, 1)).reduceByKey(lambda a, b: a + b)
b) rdd.split(" ").map(...)
c) rdd.collect()
d) rdd.countByKey()

Q.34 Which method returns only the first element of an RDD? **2**

a) head()
b) first()
c) top()
d) take(1)

Q.35 What will the following code return? **2**

```
sc.parallelize([10, 20, 30]).
map(lambda x: x + 5).
collect()
```

a) [10,20,30]
b) [15, 25, 35]
c) [5, 15, 25]
d) Error