Birla Institute of Technology and Sciences, Pilani
Comprehensive Examination 2023
Subject: Big Data Analytics (MPBA –G517) (Part-A)

Duration: 1hour 30mins
Full Marks 1x60=60

Multi-choice Questions

- *Multiple choice questions with 4 options and only 1 correct answer.*
- *Each question is of 1 marks each with no negative marking*
- *Mark the correct option in the box given below*

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
21.	22.	23.	24.	25.	26.	27.	28.	29.	30.
31.	32.	33.	34.	35.	36.	37.	38.	39.	40.
41.	42.	43.	44.	45.	46.	47.	48.	49.	50.
51.	52.	53.	54.	55.	56.	57.	58.	59.	60.

1. Which of the following is the correct way to import the pyspark module in a Python script?
a) from pyspark import *
b) import pyspark
c) import spark
d) from spark import pyspark

2. Which of the following is the correct way to create a SparkSession object in pyspark?
a) spark = pyspark.SparkSession()
b) spark = pyspark.SparkSession.builder.getOrCreate()
c) spark = SparkSession()
d) spark = SparkSession.builder.getOrCreate()

3. Which of the following is the correct way to read a CSV file into a Spark DataFrame in pyspark?
a) df = spark.read.csv("file.csv")
b) df = spark.read.format("csv").load("file.csv")
c) df = spark.read("file.csv", format="csv")
d) Both a and b are correct

4. Which of the following is the correct way to select a subset of columns from a Spark DataFrame in pyspark?
a) df.select("col1", "col2")
b) df["col1", "col2"]
c) df[["col1", "col2"]]
d) Both a and c are correct

5. Which of the following is the correct way to filter rows from a Spark DataFrame based on a condition in pyspark?
a) df.filter(df["col"] > 10)
b) df.where(df["col"] > 10)
c) df[df["col"] > 10]
d) Both a and b are correct

Q6. What docker command is used to list all the containers running on the host?
a) docker ls
b) docker ps
c) docker containers
d) docker list

Q7. What docker command is used to create a new image from a container's changes?
a) docker commit
b) docker save
c) docker build
d) docker update

Q8. What docker command is used to run a command in a running container?
a) docker exec
b) docker attach
c) docker run
d) docker start

Q9. What docker command is used to remove one or more stopped containers?
a) docker rm
b) docker rmi

c) docker prune

d) docker delete

Q10. What docker command is used to tag an image with a new name or a new version?
a) docker tag
b) docker rename
c) docker label
d) docker version

Q11. What is the output of the following code?

```
def foo(x):
if x == 0:
```

```
return 0
return x + foo(x-1)
print(foo(3))
```

A) 0
B) 3
C) 6
D) 9

Q12. What is the difference between is and == operators in Python?

A) is checks for identity, == checks for equality
B) is checks for equality, == checks for identity
C) is and == are synonyms
D) is and == have no difference

Q13. What is the output of the following code?

```
a = [1, 2, 3]
b = a
b.append(4)
print(a)
```

A) [1, 2, 3]
B) [1, 2, 3, 4]
C) [4, 1, 2, 3]
D) SyntaxError

Q14. What is the output of the following code?

```
def bar(**kwargs):
for k, v in kwargs.items():
print(k, v)
bar(a=1, b=2, c=3)
```

A) a 1 b 2 c 3
B) 1 a 2 b 3 c
C) {'a': 1, 'b': 2, 'c': 3}
D) TypeError

Q15. What is the output of the following code?

```
x = 10
def baz():
global x
x = x + 10
print(x)
baz()
print(x)
```

A) 10 10

B) 20 20
C) 10 20
D) 20 10

Q16. What is the output of the following code?

```
class Foo:
def __init__(self, x):
self.x = x
def __str__(self):
return str(self.x)
def __add__(self, other):
return Foo(self.x + other.x)
a = Foo(1)
b = Foo(2)
c = a + b
print(c)
```

A) 1
B) 2
C) 3
D) Foo(3)

Q17. What is the output of the following code?

```
import re
s = "Hello, world!"
m = re.search(r"(\w+), (\w+)!", s)
print(m.group(1))
print(m.group(2))
```

A) Hello, world!
B) Hello world
C) Hello
world
D) AttributeError

Q18. What is the output of the following code?

```
import random
random.seed(0)
print(random.randint(1, 10))
print(random.randint(1, 10))
```

A) 1 1
B) 10 10
C) 6 9
D) 9 6

Q19. What is the output of the following code?

```
def qux(n):
if n == 0:
yield "Done"
else:
yield n
yield from qux(n-1)
for x in qux(3):
print(x)
```

A) 3 2 1 0
B) 3 2 1 Done
C) Done 1 2 3
D) SyntaxError

Q20. What is the output of the following code?

```
def foo():
try:
return 1
finally:
return 2
print(foo())
```

A) 1
B) 2
C) 1 2
D) None

21. What are the benefits of using Spark over traditional MapReduce for big data processing?
A) Spark is faster, more expressive, and supports multiple languages
B) Spark is slower, less expressive, and supports only Java
C) Spark is faster, less expressive, and supports only Scala
D) Spark is slower, more expressive, and supports multiple languages

22. How does Spark achieve fault tolerance in case of node failures or data loss?
A) Spark uses checkpoints to save intermediate results to HDFS
B) Spark uses lineage graphs to reconstruct lost partitions from RDDs
C) Spark uses replication to store multiple copies of each partition
D) Spark uses backup nodes to take over the failed tasks

23. What is the basic structure of an Airflow DAG (Directed Acyclic Graph)?
A) A sequence of tasks that run in parallel
B) A set of operators that define the logic and dependencies of the workflow
C) A collection of branches that diverge and converge at different points

D) A circular path that loops through the same tasks repeatedly

24. What are the advantages of using a distributed file system like HDFS for big data processing?

A) It provides high availability, scalability, and fault tolerance.

B) It reduces the network overhead and disk I/O by moving computation to data.

C) It supports various types of data formats and schemas.

D) All of the above.

25. What are the main differences between MapReduce and Spark in terms of programming model, performance, and fault tolerance?

A) MapReduce is based on batch processing, while Spark supports both batch and stream processing.

B) MapReduce writes intermediate results to disk, while Spark caches them in memory.

C) MapReduce relies on replication for fault tolerance, while Spark uses lineage and checkpointing.

D) All of the above.

26. What are the benefits and challenges of using cloud computing platforms for big data analytics?

A) Benefits include lower cost, higher scalability, and easier access to various tools and services.

B) Challenges include data security, privacy, and governance issues.

C) Benefits include faster performance, higher reliability, and better quality of service.

D) Both A and B.

27. What are the key features of NoSQL databases that make them suitable for big data applications?

A) They are schema-less, meaning they can store and process unstructured and semi-structured data.

B) They are distributed, meaning they can scale horizontally and handle large volumes of data.

C) They are flexible, meaning they can support various data models and query languages.

D) All of the above.

28. What are some of the common tools and frameworks used for data ingestion, data cleaning, data transformation, and data visualization in a big data pipeline?

A) Data ingestion: Kafka, Flume, Sqoop, etc.

B) Data cleaning: Pandas, Spark, Hadoop, etc.

C) Data transformation: Hive, Pig, Spark, etc.

D) Data visualization: Tableau, Power BI, Matplotlib, etc.

29. What are some of the best practices and principles for designing and implementing a big data solution?

A) Define the business problem and the data requirements clearly.
B) Choose the appropriate tools and technologies based on the data characteristics and the analytical goals.
C) Implement data quality checks and validation processes throughout the pipeline.
D) All of the above.

30. Define different properties of big data?

A) Volume, velocity, variety, veracity, and value.
B) Volume, velocity, variety, veracity, and validity.
C) Volume, velocity, variety, variability, and value.
D) Volume, velocity, variety, variability, and veracity.

31. Draw the workflow framework of Docker architecture and clearly mention different components?

A) The Docker architecture consists of a client-server model, where the Docker client communicates with the Docker daemon through a REST API. The Docker daemon is responsible for building, running, and managing the containers. The Docker daemon can also communicate with other Docker daemons to form a cluster. The Docker architecture also includes a registry, which is a repository of images that can be pulled and pushed by the Docker client or daemon. The Docker architecture can be represented by the following diagram:
B) The Docker architecture consists of a peer-to-peer model, where each Docker node can act as both a client and a server. The Docker nodes can create, run, and share containers with each other. The Docker architecture also includes a swarm, which is a group of Docker nodes that can be managed as a single entity. The Docker architecture also includes a hub, which is a central repository of images that can be accessed by the Docker nodes. The Docker architecture can be represented by the following diagram:
C) The Docker architecture consists of a hybrid model, where the Docker client can communicate with either a Docker daemon or a Docker node. The Docker daemon is responsible for building, running, and managing the containers on a single host. The Docker node is responsible for creating, running, and sharing the containers with other Docker nodes. The Docker architecture also includes a registry, which is a repository of images that can be pulled and pushed by the Docker client, daemon, or node. The Docker architecture also includes a swarm, which is a group of Docker nodes that can be managed as a single entity. The Docker architecture can be represented by the following diagram:
D) None of the above.

32. Illustrate different properties of Large Language models as per your understanding?

A) Large language models are neural network models that are trained on massive amounts of text data to learn the statistical patterns and relationships of natural language. Some of the properties of large language models are:
B) Large language models are probabilistic models that are trained on a variety of text data to learn the semantic and syntactic rules of natural language. Some of the properties of large language models are:
C) Large language models are generative models that are trained on a corpus of text data to learn the latent representations and distributions of natural language. Some of the properties of large language models are:
D) All of the above.

33. Illustrate the advantages of using docker in cloud computing?

A) Docker enables faster and easier deployment of applications by packaging them into lightweight and portable containers.
B) Docker does not enhance the scalability and elasticity of applications by allowing them to run on any cloud platform that supports Docker.
C) Docker has no impact on security.
D) Docker has impact the operational workflow of the system.

34. How is big data analysis helpful in increasing revenue?

A) Big data analysis can help in increasing revenue by enabling better decision making, customer segmentation, product innovation, and market optimization.
B) Big data analysis can help in increasing revenue by reducing operational costs, improving efficiency, enhancing quality, and preventing fraud.
C) Big data analysis can help in increasing revenue by creating new business models, generating new insights, discovering new opportunities, and creating competitive advantage.
D) All of the above.

35. How would you transform unstructured data into structured data?

A) By applying data mining techniques such as clustering, classification, association, and pattern recognition.
B) By applying data extraction techniques such as parsing, tokenization, tagging, and entity recognition.
C) By applying data integration techniques such as schema mapping, record linkage, and data fusion.
D) By applying data transformation techniques such as normalization, standardization, encoding, and aggregation.

36. Explain briefly the steps involved in creating a cloud based data science solution?

A) The steps involved in creating a cloud based data science solution are:
B) The steps involved in creating a cloud based data science solution are:
C) The steps involved in creating a cloud based data science solution are:
D) None of the above.

37. What does LLM stand for?

A) Large Language Model
B) Legal Language Model
C) Linear Logic Model
D) Low Latency Model

38. What is the main challenge of training LLMs?

A) Data scarcity
B) Computational cost
C) Ethical issues
D) All of the above

Q39. What is the name of the LLM developed by OpenAI that has 175 billion parameters?

A) GPT-3
B) BERT
C) XLNet
D) Transformer-XL

Q40. What is the main advantage of LLMs over traditional NLP models?

A) They can generate natural and coherent text
B) They can perform multiple NLP tasks without fine-tuning
C) They can capture long-range dependencies and context
D) All of the above

41. What is the name of the technique that allows LLMs to generate text conditioned on a given prompt?

A) Text summarization
B) Text completion
C) Text expansion
D) Text generation

Q42. What is the name of the framework that evaluates the factual accuracy of LLM-generated text?

A) FEVER
B) GLUE
C) FactCC
D) FABRIC

Q43. What is the name of the LLM developed by Google that has 1.6 billion parameters and is pre-trained on Wikipedia and books?

A) GPT-3
B) BERT
C) XLNet
D) Transformer-XL

44. What is the name of the LLM developed by Facebook that has 9.4 billion parameters and is pre-trained on 45 languages?

A) GPT-3
B) BERT
C) XLM-R
D) RoBERTa

45. What does MLOps stand for?

A) Machine Learning Operations
B) Machine Learning Optimization
C) Machine Learning Orchestration
D) Machine Learning Ontology

Q46. What is the main goal of MLOps?

A) To automate the deployment and monitoring of ML models
B) To improve the collaboration and communication between ML engineers and data scientists
C) To reduce the gap between research and production in ML projects
D) All of the above

Q47. What are the main components of an MLOps pipeline?

A) Data ingestion, data processing, model training, model validation, model deployment, model monitoring
B) Data collection, data cleaning, data analysis, model building, model testing, model serving, model evaluation
C) Data acquisition, data transformation, model development, model verification, model delivery, model maintenance
D) All of the above are equivalent

Q48. What are some of the benefits of MLOps?

A) Faster and more reliable model deployment
B) Higher model quality and performance
C) Easier model reproducibility and traceability
D) All of the above

Q49. What are some of the challenges of MLOps?

A) Data drift and model decay
B) Model explainability and fairness
C) Model security and privacy
D) All of the above

Q50. What are some of the tools and platforms that support MLOps?

A) TensorFlow, PyTorch, Scikit-learn
B) MLflow, Kubeflow, Airflow
C) AWS, Azure, Google Cloud
D) All of the above

Q51. What is the name of the MLOps framework that is based on the DevOps principles and consists of three pillars: Continuous Integration, Continuous Delivery, and Continuous Monitoring?

A) CICD
B) CML
C) CD4ML
D) MLCI

Q52. What is the name of the MLOps framework that is based on the scientific method and consists of four phases: Plan, Execute, Monitor, and Learn?

A) PEMA
B) PDCA
C) PEMDAS
D) PMLA

Q53. What is the name of the open-source distributed processing framework that allows to process large-scale data using a cluster of commodity hardware?

A) Spark
B) Hadoop
C) Kafka
D) Storm

Q54. What are the two main components of Hadoop?

A) Hadoop Distributed File System (HDFS) and Hadoop MapReduce
B) Hadoop Streaming and Hadoop Common
C) Hadoop YARN and Hadoop Hive
D) Hadoop Pig and Hadoop ZooKeeper

Q55. What is the name of the open-source distributed computing framework that allows to perform fast and scalable data analysis and machine learning on large-scale data?

A) Spark

B) Hadoop
C) Kafka
D) Storm

Q56. What are the four main components of Spark?
A) Spark Core, Spark SQL, Spark Streaming, and Spark MLlib
B) Spark GraphX, Spark R, Spark Python, and Spark Scala
C) Spark Cluster, Spark Job, Spark Task, and Spark Stage
D) Spark Driver, Spark Master, Spark Worker, and Spark Executor

Q57. What is the name of the open-source distributed streaming platform that allows to publish and subscribe to streams of data, store them, and process them in real-time?
A) Spark
B) Hadoop
C) Kafka
D) Storm

Q58. What are the three main components of Kafka?
A) Kafka Producer, Kafka Consumer, and Kafka Broker
B) Kafka Topic, Kafka Partition, and Kafka Offset
C) Kafka Cluster, Kafka Controller, and Kafka ZooKeeper
D) All of the above

Q59. What is the name of the open-source distributed real-time computation system that allows to process streams of data in parallel and at high speed?
A) Spark
B) Hadoop
C) Kafka
D) Storm

Q60. What are the two main components of Storm?
A) Storm Nimbus and Storm Supervisor
B) Storm Spout and Storm Bolt
C) Storm Topology and Storm Stream
D) All of the above

Birla Institute of Technology and Sciences, Pilani
Comprehensive Examination 2023
Subject: Big Data Analytics (MPBA –G517) (Part-B)

Duration: 1hour 30mins
Full Marks 10x3 =30

Detailed answers questions

1. Design and explain implementation of a low-latency machine learning (LLM) based architecture for real-time data processing and inference of SQL data extraction and augmented generation? What are the benefits and challenges of using LLM over traditional batch-based approaches?

2. What are the key components and features of a cloud based recommender engine system? How would you ensure scalability, reliability, and security of the system? What are some of the common algorithms and techniques used for recommendation and personalization?

3. How would you approach the problem of news analysis and clustering from multiple sources and languages? What are the steps and tools involved in data collection, preprocessing, feature extraction, clustering, and visualization? How would you evaluate the quality and relevance of the clusters?