### ****Big Data Concepts****

Which of the following is NOT a source of Big Data?  
a) Social Media  
b) Scientific Experiments  
c) Personal Diaries  
d) Business Transactions  
**Answer**: c) Personal Diaries

What is the primary focus of data curation in repositories?  
a) Analyzing big data trends  
b) Organizing and maintaining data for reuse  
c) Selling data to third parties  
d) Replacing outdated systems  
**Answer**: b) Organizing and maintaining data for reuse

The term "Open Exit" in the data life cycle refers to:  
a) Starting a new data project  
b) Archiving data at the end of its life  
c) Deleting old datasets  
d) Sharing data publicly  
**Answer**: b) Archiving data at the end of its life

### ****HDFS Basics****

What does HDFS stand for?  
a) Highly Distributed File Storage  
b) Hadoop Distributed File System  
c) High-Density File Storage  
d) Hadoop Data File Server  
**Answer**: b) Hadoop Distributed File System

Which daemon is responsible for managing the metadata in HDFS?  
a) Data Node  
b) Name Node  
c) Secondary Name Node  
d) Task Tracker  
**Answer**: b) Name Node

How does HDFS achieve fault tolerance?  
a) Through replication  
b) Using a backup server  
c) By storing data centrally  
d) Through encryption  
**Answer**: a) Through replication

The replication factor in HDFS ensures:  
a) High-speed data processing  
b) Fault tolerance and high availability  
c) Faster network communication  
d) Distributed computation  
**Answer**: b) Fault tolerance and high availability

### ****HDFS Operations****

What is the role of the Secondary Name Node in HDFS?  
a) It processes client requests  
b) It takes over if the Name Node fails  
c) It merges the Name Node’s edit logs  
d) It handles block placement  
**Answer**: c) It merges the Name Node’s edit logs

Which HDFS command lists all files in a directory?  
a) hdfs ls  
b) hadoop list  
c) hdfs dfs -ls  
d) hadoop fs -list  
**Answer**: c) hdfs dfs -ls

### ****Hadoop MapReduce****

In the MapReduce paradigm, the role of the Mapper is to:  
a) Combine intermediate data  
b) Split the input data into smaller chunks  
c) Generate key-value pairs  
d) Sort the output data  
**Answer**: c) Generate key-value pairs

Which of these is NOT a MapReduce daemon?  
a) Job Tracker  
b) Task Tracker  
c) Resource Manager  
d) Name Node  
**Answer**: d) Name Node

What is the function of a Partitioner in MapReduce?  
a) Distributes data to reducers  
b) Compresses the data  
c) Generates input splits  
d) Executes the mapper  
**Answer**: a) Distributes data to reducers

What is the primary input format in Hadoop MapReduce?  
a) Text Input Format  
b) Binary Input Format  
c) JSON Input Format  
d) XML Input Format  
**Answer**: a) Text Input Format

### ****Hadoop Architecture****

Which is NOT a core component of Hadoop?  
a) HDFS  
b) MapReduce  
c) Yarn  
d) MongoDB  
**Answer**: d) MongoDB

What is the default replication factor in HDFS?  
a) 1  
b) 2  
c) 3  
d) 4  
**Answer**: c) 3

### ****Advanced Hadoop Concepts****

Distributed cache in Hadoop is used for:  
a) Temporary storage  
b) Sharing small files between nodes  
c) Managing task distribution  
d) Replicating data  
**Answer**: b) Sharing small files between nodes

Hadoop ETL tools are primarily used for:  
a) Database indexing  
b) Extracting, Transforming, and Loading data  
c) Creating MapReduce programs  
d) Managing Hadoop clusters  
**Answer**: b) Extracting, Transforming, and Loading data

### ****Hands-On Tasks****

Which command uploads a file to HDFS?  
a) hadoop fs -put  
b) hdfs upload  
c) hdfs add  
d) hadoop addfile  
**Answer**: a) hadoop fs -put

To check cluster status, you use the command:  
a) hdfs dfsadmin -report  
b) hdfs cluster -status  
c) hadoop check  
d) hadoop report  
**Answer**: a) hdfs dfsadmin -report

### ****Security and Maintenance****

Which of the following ensures secure access to HDFS?  
a) Encryption  
b) Rack Awareness  
c) Kerberos Authentication  
d) Data Pipelining  
**Answer**: c) Kerberos Authentication

What is the purpose of Hadoop benchmarks?  
a) Measuring cluster performance  
b) Enhancing data security  
c) Checking replication consistency  
d) Validating job execution  
**Answer**: a) Measuring cluster performance

### ****Programming with Hadoop****

What are Hadoop Data Types used for?  
a) Storing metadata  
b) Representing keys and values in MapReduce  
c) Defining user permissions  
d) Managing cluster configurations  
**Answer**: b) Representing keys and values in MapReduce

Which of the following is an example of a MapReduce Java data type?  
a) Text  
b) String  
c) Binary  
d) FloatArray  
**Answer**: a) Text

### ****Miscellaneous****

1. What is the role of Combiners in Hadoop?  
   a) Reduce the amount of data transferred to reducers  
   b) Combine input splits  
   c) Assign tasks to reducers  
   d) Manage output formatting  
   **Answer**: a) Reduce the amount of data transferred to reducers

### ****HDFS Operations and Commands****

Which command in HDFS is used to remove a file?  
a) hdfs dfs -delete  
b) hadoop fs -rm  
c) hdfs remove  
d) hadoop dfs -erase  
**Answer**: b) hadoop fs -rm

The hdfs dfs -copyFromLocal command is used for:  
a) Copying files from HDFS to the local filesystem  
b) Uploading files from the local filesystem to HDFS  
c) Copying files between HDFS directories  
d) Backing up files from the local filesystem  
**Answer**: b) Uploading files from the local filesystem to HDFS

### ****Hadoop ETL****

What is the primary advantage of using Hadoop ETL tools?  
a) Simplifying data extraction and transformation at scale  
b) Improving cluster configuration  
c) Reducing the number of Hadoop nodes  
d) Managing cluster security  
**Answer**: a) Simplifying data extraction and transformation at scale

In Hadoop ETL, which of the following represents a data extraction step?  
a) Loading data into HDFS  
b) Parsing log files into structured data  
c) Transforming data into a new format  
d) Aggregating data for analysis  
**Answer**: b) Parsing log files into structured data

### ****MapReduce Programming****

Which MapReduce feature is used to count the number of occurrences of specific events during execution?  
a) Counters  
b) Partitioners  
c) Combiners  
d) Reducers  
**Answer**: a) Counters

Which of the following is a valid use case for the Distributed Cache in MapReduce?  
a) Distributing configuration files to all nodes  
b) Executing parallel MapReduce jobs  
c) Writing output data to multiple locations  
d) Reducing input split size  
**Answer**: a) Distributing configuration files to all nodes

#### ****Introduction to HBase****

HBase is primarily designed for:  
a) Relational database management  
b) Processing real-time analytics  
c) Handling large-scale, distributed, non-relational data  
d) Transactional processing  
**Answer**: c) Handling large-scale, distributed, non-relational data

Which of the following is a core feature of HBase?  
a) Structured data storage  
b) Random real-time read/write access to Big Data  
c) Full ACID compliance  
d) Relational data modeling  
**Answer**: b) Random real-time read/write access to Big Data

HBase is built on top of which of the following systems?  
a) Hadoop Distributed File System (HDFS)  
b) Apache Spark  
c) MongoDB  
d) MySQL  
**Answer**: a) Hadoop Distributed File System (HDFS)

What is the default file format used in HBase?  
a) JSON  
b) XML  
c) HFile  
d) Parquet  
**Answer**: c) HFile

#### ****HBase Architecture****

Which HBase component is responsible for managing region servers?  
a) Name Node  
b) Master Server  
c) Zookeeper  
d) Data Node  
**Answer**: b) Master Server

What is the role of Zookeeper in HBase?  
a) Storing data  
b) Managing schema updates  
c) Coordinating distributed systems  
d) Handling data replication  
**Answer**: c) Coordinating distributed systems

Regions in HBase are:  
a) Fixed-size partitions of tables  
b) Variable-sized partitions of tables  
c) Stored in Zookeeper  
d) Managed by Data Nodes  
**Answer**: b) Variable-sized partitions of tables

What is the default block size for HBase tables?  
a) 32 KB  
b) 64 KB  
c) 64 MB  
d) 128 MB  
**Answer**: d) 128 MB

#### ****Installation and HBase Security****

Which of the following is NOT required to set up HBase?  
a) Hadoop  
b) Zookeeper  
c) PostgreSQL  
d) Java  
**Answer**: c) PostgreSQL

In HBase security, Kerberos is used for:  
a) Data replication  
b) Authentication  
c) File compression  
d) Load balancing  
**Answer**: b) Authentication

#### ****HBase Operations and Shell Commands****

Which command is used to create a table in HBase shell?  
a) create\_table  
b) table\_create  
c) create  
d) new\_table  
**Answer**: c) create

What does the HBase scan command do?  
a) Deletes a table  
b) Reads data sequentially from a table  
c) Writes data to a table  
d) Displays table schema  
**Answer**: b) Reads data sequentially from a table

The HBase shell command to count rows in a table is:  
a) count  
b) row\_count  
c) table\_count  
d) get\_row\_count  
**Answer**: a) count

To truncate a table in HBase, which of the following commands is used?  
a) delete  
b) truncate  
c) drop\_table  
d) remove\_table  
**Answer**: b) truncate

#### ****HBase Java Client and API****

In HBase, CRUD operations are implemented using which Java class?  
a) HBaseAdmin  
b) HTable  
c) HFile  
d) HRegion  
**Answer**: b) HTable

The method used to add a row to an HBase table using Java API is:  
a) putRow  
b) insert  
c) add  
d) put  
**Answer**: d) put

The Java API to delete a row from an HBase table is:  
a) remove  
b) deleteRow  
c) delete  
d) removeRow  
**Answer**: c) delete

Which API is used for administrative tasks in HBase?  
a) HBaseClient  
b) AdminAPI  
c) HBaseAdmin  
d) MasterAPI  
**Answer**: c) HBaseAdmin

#### ****HBase Lab Assignments****

What does the following HBase shell command do?  
list  
a) Lists all the regions  
b) Lists all tables in the database  
c) Lists the columns in a table  
d) Lists all commands available in the shell  
**Answer**: b) Lists all tables in the database

What does the Java method scan() return when used in HBase?  
a) A single row from the table  
b) All matching rows in the table  
c) The schema of the table  
d) The total number of rows in the table  
**Answer**: b) All matching rows in the table

### ****Intermediate-Level MCQs on Hive****

#### ****Introduction to Hive and Architecture****

Hive is best described as:  
a) A relational database management system  
b) A data warehouse system on top of Hadoop  
c) A NoSQL database  
d) A distributed file system  
**Answer**: b) A data warehouse system on top of Hadoop

Which component is NOT part of the Hive architecture?  
a) Metastore  
b) Execution Engine  
c) Job Tracker  
d) Driver  
**Answer**: c) Job Tracker

Hive queries are converted into:  
a) SQL queries  
b) MapReduce jobs  
c) Spark transformations  
d) NoSQL operations  
**Answer**: b) MapReduce jobs

How does Hive differ from traditional RDBMS?  
a) It supports ACID transactions  
b) It processes data in real-time  
c) It is optimized for batch processing of large datasets  
d) It uses a fixed schema  
**Answer**: c) It is optimized for batch processing of large datasets

#### ****Hive Query Language (HQL) Basics****

Which of the following is NOT a valid Hive data type?  
a) STRUCT  
b) FLOAT  
c) JSON  
d) MAP  
**Answer**: c) JSON

What is the purpose of the LIKE operator in Hive?  
a) To perform arithmetic operations  
b) To perform pattern matching  
c) To define partitions  
d) To import data  
**Answer**: b) To perform pattern matching

Which keyword is used to remove duplicates in a Hive query?  
a) UNIQUE  
b) DISTINCT  
c) DELETE  
d) EXCEPT  
**Answer**: b) DISTINCT

In Hive, what does the OVERWRITE option do when loading data into a table?  
a) Appends the data  
b) Deletes the table schema  
c) Replaces existing data in the table  
d) Merges the data  
**Answer**: c) Replaces existing data in the table

#### ****Tables and Partitions****

Which of the following is TRUE about Managed Tables in Hive?  
a) The table data is stored externally  
b) Dropping the table deletes both the table and data  
c) Data is stored in a NoSQL format  
d) Managed Tables do not support partitions  
**Answer**: b) Dropping the table deletes both the table and data

What is the main purpose of partitions in Hive?  
a) To create indexes  
b) To improve query performance  
c) To store metadata  
d) To reduce data redundancy  
**Answer**: b) To improve query performance

Bucketing in Hive is primarily used to:  
a) Replicate data across nodes  
b) Perform better joins and sampling  
c) Store unstructured data  
d) Index the table columns  
**Answer**: b) Perform better joins and sampling

How is data organized when stored in ORC (Optimized Row Columnar) format?  
a) Row-wise  
b) Column-wise  
c) Key-value pairs  
d) JSON objects  
**Answer**: b) Column-wise

#### ****Loading and Manipulating Data****

The command to load local data into a Hive table is:  
a) LOAD DATA INPATH  
b) LOAD DATA LOCAL INPATH  
c) INSERT INTO  
d) COPY DATA  
**Answer**: b) LOAD DATA LOCAL INPATH

Which of the following allows appending data to an existing table in Hive?  
a) OVERWRITE  
b) INSERT INTO  
c) REPLACE DATA  
d) CREATE APPEND  
**Answer**: b) INSERT INTO

To delete the schema and all its data, you use:  
a) DROP TABLE  
b) DELETE TABLE  
c) TRUNCATE TABLE  
d) REMOVE TABLE  
**Answer**: a) DROP TABLE

#### ****Querying Data****

Which of the following is NOT a type of join supported by Hive?  
a) Inner Join  
b) Left Outer Join  
c) Full Outer Join  
d) Cross Join  
**Answer**: d) Cross Join

What does the following HQL query do?  
SELECT COUNT(\*) FROM table\_name;  
a) Returns the total number of columns  
b) Counts the rows in the table  
c) Removes duplicates from the table  
d) Deletes rows from the table  
**Answer**: b) Counts the rows in the table

In Hive, a Subquery is:  
a) A query inside another query  
b) A type of table join  
c) A query to create indexes  
d) Used for data import  
**Answer**: a) A query inside another query

#### ****Optimization Techniques****

Map-side joins are used to:  
a) Reduce the amount of data transferred between nodes  
b) Avoid the creation of reducers  
c) Load data into partitions  
d) Create indexes on tables  
**Answer**: a) Reduce the amount of data transferred between nodes

Distributed Cache in Hive is used to:  
a) Store intermediate query results  
b) Share small read-only files across nodes  
c) Perform real-time data analysis  
d) Backup metadata  
**Answer**: b) Share small read-only files across nodes

#### ****UDFs and Custom Scripts****

What does UDF stand for in Hive?  
a) Unified Data Format  
b) User-Defined Function  
c) Universal Data Framework  
d) User Data Fetcher  
**Answer**: b) User-Defined Function

Hive UDFs are used for:  
a) Defining schema  
b) Performing custom operations on data  
c) Loading data into tables  
d) Managing table partitions  
**Answer**: b) Performing custom operations on data

#### ****Views and Advanced Operations****

A View in Hive is:  
a) A physical storage of data  
b) A virtual table generated by a query  
c) A backup of a table  
d) A schema definition  
**Answer**: b) A virtual table generated by a query

Which of the following commands is used to create a view in Hive?  
a) CREATE TABLE  
b) CREATE VIEW  
c) DEFINE VIEW  
d) CREATE SCHEMA  
**Answer**: b) CREATE VIEW

#### ****Lab Assignments****

To create an external table in Hive, which keyword is used?  
a) MANAGED  
b) EXTERNAL  
c) UNMANAGED  
d) REMOTE  
**Answer**: b) EXTERNAL

The command to truncate a Hive table is:  
a) DELETE  
b) REMOVE  
c) TRUNCATE  
d) DROP  
**Answer**: c) TRUNCATE

#### ****Miscellaneous****

Which of the following is an advantage of using Hive?  
a) Real-time data updates  
b) High latency for small datasets  
c) Integration with Hadoop for large-scale data processing  
d) Fixed schema support  
**Answer**: c) Integration with Hadoop for large-scale data processing

The .q file extension in Hive is associated with:  
a) Metadata files  
b) HQL script files  
c) Configuration files  
d) Log files  
**Answer**: b) HQL script files

Which storage format in Hive is best suited for faster query performance?  
a) TextFile  
b) SequenceFile  
c) ORC  
d) Avro  
**Answer**: c) ORC

To optimize queries, which of the following can be used in Hive?  
a) Partitioning  
b) Indexing  
c) Bucketing  
d) All of the above  
**Answer**: d) All of the above

### ****Intermediate-Level MCQs on Data Warehousing, Data Lakes, and Apache Spark****

#### ****Data Warehousing and Data Lakes****

A Data Warehouse is optimized for:  
a) Real-time data processing  
b) Transactional operations  
c) Analytical querying and reporting  
d) Storing unstructured data  
**Answer**: c) Analytical querying and reporting

What is the main purpose of a Data Lake?  
a) To store structured and semi-structured data only  
b) To store all types of raw data for future processing  
c) To act as a replacement for relational databases  
d) To store transactional data exclusively  
**Answer**: b) To store all types of raw data for future processing

Which of the following is a key characteristic of a Data Warehouse?  
a) Schema-on-read  
b) Schema-on-write  
c) No schema required  
d) Data stored as objects  
**Answer**: b) Schema-on-write

In an ETL data pipeline, the "E" stands for:  
a) Evaluate  
b) Extract  
c) Execute  
d) Enrich  
**Answer**: b) Extract

#### ****ETL vs. ELT****

What is the primary difference between ETL and ELT?  
a) ELT processes data before loading, while ETL processes data after loading  
b) ETL uses in-database transformations, while ELT extracts raw data  
c) ETL performs transformations before loading, while ELT transforms data after loading  
d) ETL is faster than ELT for all use cases  
**Answer**: c) ETL performs transformations before loading, while ELT transforms data after loading

ELT is most commonly used with:  
a) Data Warehouses  
b) Data Lakes  
c) NoSQL Databases  
d) Flat files  
**Answer**: b) Data Lakes

#### ****Airflow/Informatica****

In Apache Airflow, a DAG represents:  
a) Data Analytics Graph  
b) Directed Acyclic Graph  
c) Data Arrangement Guide  
d) Distributed Analytics Graph  
**Answer**: b) Directed Acyclic Graph

Informatica is primarily used for:  
a) Streaming data processing  
b) Extracting, Transforming, and Loading data  
c) Managing machine learning workflows  
d) Managing distributed databases  
**Answer**: b) Extracting, Transforming, and Loading data

Which of the following is a feature of Airflow?  
a) Managing structured data storage  
b) Workflow orchestration and scheduling  
c) Real-time data streaming  
d) Cloud-based data storage  
**Answer**: b) Workflow orchestration and scheduling

In Airflow, what is an operator?  
a) A reusable task that performs a specific function in the pipeline  
b) A data transformation tool  
c) A type of storage for pipeline results  
d) A log aggregation service  
**Answer**: a) A reusable task that performs a specific function in the pipeline

#### ****Apache Spark Basics****

Apache Spark is a framework for:  
a) Real-time transactional processing  
b) Large-scale distributed data processing  
c) Managing relational databases  
d) Creating NoSQL data stores  
**Answer**: b) Large-scale distributed data processing

What is an RDD in Spark?  
a) Random Distributed Dataset  
b) Resilient Distributed Dataset  
c) Reliable Distributed Data  
d) Real-time Data Distribution  
**Answer**: b) Resilient Distributed Dataset

Which of the following provides better optimization and performance in Spark?  
a) RDD  
b) DataFrames  
c) Key-Value Pairs  
d) Hadoop Streaming  
**Answer**: b) DataFrames

#### ****Spark Operations and Streaming****

In Spark, what does the persist() method do?  
a) Deletes an RDD from memory  
b) Saves an RDD to memory or disk for reuse  
c) Optimizes shuffle operations  
d) Removes duplicates from an RDD  
**Answer**: b) Saves an RDD to memory or disk for reuse

Structured Streaming in Spark is used for:  
a) Querying batch data  
b) Processing real-time streaming data  
c) Generating random datasets  
d) Creating NoSQL databases  
**Answer**: b) Processing real-time streaming data

The main advantage of DataFrames over RDDs is:  
a) Support for schema-less data  
b) Compatibility with SQL-like queries  
c) Ability to store unstructured data  
d) Direct integration with NoSQL databases  
**Answer**: b) Compatibility with SQL-like queries

The reduceByKey operation in Spark is applied to:  
a) DataFrames only  
b) Key-Value RDDs  
c) External Datasets  
d) Non-pair RDDs  
**Answer**: b) Key-Value RDDs

#### ****Spark Integrations and Deployment****

Spark can run without Hadoop by:  
a) Using local mode or Mesos/YARN as resource managers  
b) Using only DataFrames and Datasets  
c) Disabling distributed processing  
d) Converting all data to JSON format  
**Answer**: a) Using local mode or Mesos/YARN as resource managers

What does JDBC stand for in Spark?  
a) Java Database Connectivity  
b) Joint Data Backup and Connection  
c) Java Distributed Cluster  
d) JSON Data Backup Client  
**Answer**: a) Java Database Connectivity

In Spark JDBC connectivity, the driver class is specified to:  
a) Read and write data to external databases  
b) Process real-time data streams  
c) Convert JSON data to SQL tables  
d) Create NoSQL connections  
**Answer**: a) Read and write data to external databases

#### ****Data Preprocessing and EDA****

What is the purpose of data preprocessing?  
a) Transform raw data into a format suitable for analysis  
b) Store raw data in Data Warehouses  
c) Process queries in real-time  
d) Convert structured data to unstructured data  
**Answer**: a) Transform raw data into a format suitable for analysis

In Exploratory Data Analysis (EDA), the primary goal is to:  
a) Clean the dataset  
b) Summarize the main characteristics of the data  
c) Build machine learning models  
d) Deploy the dataset to a cluster  
**Answer**: b) Summarize the main characteristics of the data

#### ****Programming with Spark****

What is the function of the map() operation in Spark?  
a) Transforms each element of an RDD  
b) Groups elements by key  
c) Sorts the RDD  
d) Removes duplicates from the RDD  
**Answer**: a) Transforms each element of an RDD

In Spark, shared variables are used to:  
a) Create redundant copies of data  
b) Share data efficiently across tasks  
c) Store metadata about RDDs  
d) Partition data for processing  
**Answer**: b) Share data efficiently across tasks

#### ****Custom Pipelines and MapReduce****

A MapReduce operation consists of:  
a) Extract, Transform, and Load steps  
b) Map, Shuffle, and Reduce steps  
c) Input, Processing, and Output steps  
d) Partition, Key, and Value steps  
**Answer**: b) Map, Shuffle, and Reduce steps

In Spark, the driver program is responsible for:  
a) Distributing data to worker nodes  
b) Coordinating and managing the entire Spark job  
c) Processing data on individual nodes  
d) Managing the storage of data on HDFS  
**Answer**: b) Coordinating and managing the entire Spark job

The command to initialize a Spark session in Python is:  
a) start\_spark\_session()  
b) spark.init()  
c) SparkSession.builder()  
d) create\_session()  
**Answer**: c) SparkSession.builder()

#### ****Miscellaneous****

Which of the following supports both batch and streaming data in Spark?  
a) RDD  
b) DataFrame  
c) Dataset  
d) Structured Streaming  
**Answer**: d) Structured Streaming

Airflow DAGs are written in:  
a) SQL  
b) Python  
c) Java  
d) Scala  
**Answer**: b) Python

Shuffle operations in Spark involve:  
a) Redistributing data across partitions  
b) Removing duplicates from RDDs  
c) Sorting data within partitions  
d) Adding metadata to RDDs  
**Answer**: a) Redistributing data across partitions

#### ****Data Warehousing and Data Lakes****

Which of the following is a key advantage of a Data Lake over a Data Warehouse?  
a) Faster query performance for structured data  
b) Ability to store raw, unprocessed data of all types  
c) Stronger support for relational queries  
d) Better data governance capabilities  
**Answer**: b) Ability to store raw, unprocessed data of all types

What is the primary role of a staging area in a Data Warehouse?  
a) To clean and preprocess data before loading  
b) To store processed data for analytics  
c) To archive old data  
d) To manage unstructured data storage  
**Answer**: a) To clean and preprocess data before loading

#### ****ETL vs. ELT****

1. ELT processes are generally faster than ETL when:  
   a) The transformations are simple  
   b) The target system has strong computational power  
   c) The data is structured  
   d) The source system has better processing power  
   **Answer**: b) The target system has strong computational power

#### ****Airflow/Informatica****

In Airflow, which component is responsible for executing tasks?  
a) Scheduler  
b) Worker  
c) DAG  
d) Metadata Database  
**Answer**: b) Worker

What is the primary purpose of Informatica’s transformation tools?  
a) To automate task scheduling  
b) To apply business rules to data  
c) To manage data governance policies  
d) To schedule batch workflows  
**Answer**: b) To apply business rules to data

#### ****Apache Spark Basics****

1. What is the primary purpose of the SparkContext object?  
   a) To define DataFrame schemas  
   b) To coordinate all Spark operations  
   c) To initialize cluster nodes  
   d) To manage distributed variables  
   **Answer**: b) To coordinate all Spark operations

#### ****Spark Operations and Streaming****

1. Which method is used to perform an action that returns the first element of an RDD?  
   a) head()  
   b) first()  
   c) take(1)  
   d) collect()  
   **Answer**: b) first()

#### ****Programming with Spark****

In Spark, which method is used to broadcast a variable across all nodes?  
a) spark.broadcast()  
b) SparkContext.broadcast()  
c) SparkSession.broadcast()  
d) RDD.broadcast()  
**Answer**: b) SparkContext.broadcast()

The filter() transformation in Spark:  
a) Combines multiple RDDs  
b) Removes duplicates from an RDD  
c) Returns a new RDD containing only elements that satisfy a condition  
d) Splits an RDD into partitions  
**Answer**: c) Returns a new RDD containing only elements that satisfy a condition

#### ****Miscellaneous****

1. What does EDA primarily help identify in a dataset?  
   a) Missing data, trends, and outliers  
   b) Optimal data storage formats  
   c) Database normalization issues  
   d) Advanced machine learning algorithms  
   **Answer**: a) Missing data, trends, and outliers

### ****Introduction to Kafka****

Kafka is primarily designed for:  
a) Data storage and analytics  
b) Real-time data streaming and messaging  
c) Machine learning model deployment  
d) Database transactions  
**Answer**: b) Real-time data streaming and messaging

In Kafka, a "Topic" is:  
a) A unit of storage  
b) A unique identifier for a message  
c) A category to which messages are published  
d) A data processing unit  
**Answer**: c) A category to which messages are published

What is the purpose of a Kafka partition?  
a) To store metadata  
b) To distribute messages across brokers for scalability  
c) To process data in batch mode  
d) To store consumer information  
**Answer**: b) To distribute messages across brokers for scalability

The Kafka Producer API is used to:  
a) Read messages from topics  
b) Publish messages to topics  
c) Monitor Kafka brokers  
d) Replicate data across brokers  
**Answer**: b) Publish messages to topics

### ****Working with Kafka Using Spark****

Spark connects to Kafka using:  
a) JDBC Driver  
b) Kafka Direct Stream  
c) Kafka Connect API  
d) HDFS API  
**Answer**: b) Kafka Direct Stream

Which of the following is an advantage of using Spark with Kafka?  
a) Real-time data processing with fault tolerance  
b) High availability of data storage  
c) Faster data ingestion  
d) Reduced need for cluster resources  
**Answer**: a) Real-time data processing with fault tolerance

### ****Spark Streaming Architecture and APIs****

Spark Streaming uses which type of processing model?  
a) Batch processing  
b) Micro-batch processing  
c) Stream-first processing  
d) Event-driven processing  
**Answer**: b) Micro-batch processing

In Spark Streaming, what is a "DStream"?  
a) A distributed RDD for static data  
b) A continuous stream of RDDs for streaming data  
c) A data partitioning method  
d) A method for parallel processing  
**Answer**: b) A continuous stream of RDDs for streaming data

The Spark Streaming API provides:  
a) Direct access to SQL-like queries  
b) Functions for processing real-time data streams  
c) Tools for managing database connections  
d) Algorithms for machine learning  
**Answer**: b) Functions for processing real-time data streams

### ****Building Stream Processing Applications with Spark****

Which operation is NOT supported in Spark Streaming?  
a) Windowed operations  
b) Map transformations  
c) RDD persistence  
d) SQL table creation  
**Answer**: d) SQL table creation

To persist a Spark Streaming application’s state, which feature is used?  
a) Checkpointing  
b) Caching  
c) Partitioning  
d) Buffering  
**Answer**: a) Checkpointing

What is a primary use case of Spark Streaming with Kafka?  
a) Batch data analysis  
b) Real-time event monitoring  
c) Data storage in HDFS  
d) Predictive modeling  
**Answer**: b) Real-time event monitoring

### ****Setting Up Kafka Producer and Consumer****

Which of the following commands creates a Kafka topic?  
a) kafka-topic  
b) kafka-topics.sh  
c) create-topic.sh  
d) kafka-produce  
**Answer**: b) kafka-topics.sh

The Kafka Consumer API is responsible for:  
a) Creating topics  
b) Reading messages from topics  
c) Replicating data across brokers  
d) Sending metadata to producers  
**Answer**: b) Reading messages from topics

In Kafka, consumer groups allow:  
a) Load balancing of message consumption  
b) Partitioning of data  
c) Replication of data  
d) Topic creation  
**Answer**: a) Load balancing of message consumption

### ****Spark SQL****

Spark SQL is used to:  
a) Write MapReduce programs  
b) Query structured data using SQL-like syntax  
c) Process real-time streams  
d) Manage Spark jobs  
**Answer**: b) Query structured data using SQL-like syntax

What is the primary abstraction used in Spark SQL?  
a) RDDs  
b) DStreams  
c) DataFrames  
d) Partitions  
**Answer**: c) DataFrames

Which command in Spark SQL creates a table?  
a) CREATE  
b) CREATE TABLE  
c) NEW TABLE  
d) DEFINE TABLE  
**Answer**: b) CREATE TABLE

### ****Spark MLlib and Predictive Analysis****

What is Spark MLlib?  
a) A library for real-time data ingestion  
b) A distributed machine learning library  
c) A tool for managing Spark jobs  
d) A data visualization library  
**Answer**: b) A distributed machine learning library

Which method is used to train a linear regression model in Spark MLlib?  
a) LinearRegression.fit()  
b) trainLinearModel()  
c) LinearRegression.train()  
d) train()  
**Answer**: a) LinearRegression.fit()

Predictive analytics in Spark MLlib typically uses which type of data?  
a) Structured data  
b) Semi-structured data  
c) Unstructured data  
d) Time-series data  
**Answer**: a) Structured data

### ****Deep Learning with Spark****

Deep learning in Spark can be implemented using:  
a) TensorFlowOnSpark  
b) SparkCoreDeep  
c) HDFSDeepLearning  
d) KafkaDeepModel  
**Answer**: a) TensorFlowOnSpark

What is the advantage of Spark in deep learning?  
a) Real-time predictions  
b) Parallelized model training at scale  
c) Data visualization tools  
d) Prebuilt deep learning models  
**Answer**: b) Parallelized model training at scale

### ****Database Connectivity and Workflows****

Which library is commonly used for JDBC connections in Spark?  
a) org.apache.spark.jdbc  
b) org.apache.spark.sql  
c) org.apache.spark.data  
d) org.apache.spark.connect  
**Answer**: b) org.apache.spark.sql

In Airflow, the BashOperator is used to:  
a) Execute bash commands in a workflow  
b) Manage Python tasks  
c) Create Kafka producers  
d) Perform SQL queries  
**Answer**: a) Execute bash commands in a workflow

Which Python operator in Airflow is used for ingesting data?  
a) PythonOperator  
b) BashOperator  
c) DataTransferOperator  
d) TransferPythonOperator  
**Answer**: a) PythonOperator

### ****Miscellaneous****

Spark Streaming integrates with HDFS to:  
a) Store streaming data as files  
b) Perform in-memory computations  
c) Query real-time data  
d) Apply machine learning models  
**Answer**: a) Store streaming data as files

Kafka Connect is used for:  
a) Building custom producers and consumers  
b) Connecting Kafka to external data systems  
c) Managing partition replication  
d) Monitoring Kafka brokers  
**Answer**: b) Connecting Kafka to external data systems

In a Spark Streaming application, sliding window operations allow:  
a) Processing data across overlapping time intervals  
b) Storing data in HDFS  
c) Querying static datasets  
d) Monitoring Spark jobs  
**Answer**: a) Processing data across overlapping time intervals

Which Spark MLlib method is used to normalize data?  
a) Normalizer.transform()  
b) normalize()  
c) DataNormalizer.apply()  
d) DataFrame.normal()  
**Answer**: a) Normalizer.transform()