Domain 1: Data Engineering – Practice MCQs

1. Which AWS storage service is most suitable for storing large-scale, unstructured data that needs to be accessed by multiple ML processes?

A. Amazon EFS

B. Amazon S3

C. Amazon EBS

D. Amazon RDS

2. You have a daily batch job that reads data from Amazon S3, transforms it using PySpark, and writes the output back to S3. Which AWS service is best suited for orchestrating this workflow?

A. AWS Step Functions

B. AWS CloudFormation

C. Amazon Managed Streaming for Apache Kafka

D. Amazon Simple Queue Service (SQS)

3. A startup wants to store clickstream data that is continuously generated by millions of users, and then run near-real-time analytics on this data. Which AWS service combination is most appropriate?

A. Amazon Kinesis Data Streams and AWS Glue

B. Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics (Managed Service for Apache Flink)

C. Amazon EMR and Amazon Athena

D. Amazon EC2 and Amazon EBS

4. Your team needs to ingest and process real-time sensor data from thousands of IoT devices. After minimal processing, data should be stored for long-term analytics. Which option describes the best approach?

A. Collect data with Amazon Kinesis Data Firehose and deliver it to Amazon S3

B. Use Amazon Athena to directly query data from devices

C. Create an AWS Batch job that polls devices for data every minute

D. Use AWS Glue triggers to pull data from each device’s local database

5. Which of the following is not a typical use case for Amazon EMR in a data engineering pipeline?

A. Running large-scale batch processing with Apache Spark

B. Streaming real-time data from IoT devices at high throughput

C. Performing ETL jobs using Hadoop ecosystem tools

D. Ad hoc querying of large data sets with Hive

6. In a streaming data pipeline, which AWS service can you use to reliably capture, transform, and load streaming data into data stores such as Amazon S3 or Amazon Redshift?

A. AWS Glue

B. Amazon Kinesis Data Firehose

C. Amazon DynamoDB Streams

D. Amazon Kinesis Data Streams (directly)

7. You need to schedule a nightly ETL job that uses AWS Glue to read data from an on-premises database and write it to Amazon S3 in Parquet format. Which AWS feature or service should you use to schedule this job?

A. AWS CloudTrail

B. AWS Lambda

C. AWS Glue triggers

D. AWS DataSync

8. When deciding between Amazon EFS and Amazon EBS for storing data used in machine learning training, which factor is most important to consider?

A. EFS has higher maximum throughput than EBS

B. EBS is a regional service whereas EFS is tied to a single Availability Zone

C. EBS can be attached to multiple EC2 instances simultaneously

D. EFS provides a shared file system that can be accessed by multiple instances

9. Which AWS service is most suitable for setting up and running a Hadoop or Spark cluster for large-scale data processing?

A. Amazon Athena

B. Amazon EMR

C. AWS Glue

D. Amazon Redshift

10. You want to automate the extraction of metadata from a CSV dataset in Amazon S3 and create a data catalog for future use. Which service should you consider first?

A. AWS Glue

B. Amazon EMR

C. Amazon RDS

D. Amazon Elasticsearch Service

11. A machine learning team needs to store structured transactional data with high availability and low latency for real-time updates. Which AWS service would be the best fit?

A. Amazon DynamoDB

B. Amazon S3

C. Amazon EBS

D. Amazon Redshift

12. You need to process a dataset with both batch and streaming requirements. The batch jobs run monthly, but there is also a real-time analytics requirement for a subset of the data. Which combination of services is most appropriate?

A. Amazon EMR for batch, Amazon Kinesis Data Analytics for streaming

B. AWS Batch for batch, Amazon QuickSight for streaming

C. AWS Glue for batch, Amazon DynamoDB Streams for streaming

D. Amazon Athena for batch, AWS Lambda for streaming

13. In a data lake architecture on AWS, which component typically serves as the primary storage layer?

A. Amazon Redshift

B. Amazon S3

C. AWS Glue Data Catalog

D. Amazon Elasticsearch Service

14. You are planning to use Amazon Kinesis Data Streams to handle real-time ingestion of financial transactions. Which of the following is a primary advantage of Kinesis Data Streams in this scenario?

A. Automatic data encryption at rest

B. Seamless integration with on-premises mainframes

C. Ability to replay and reprocess events

D. On-demand schema inference for structured data

15. A data scientist wants to run ad hoc SQL queries on massive datasets stored in Amazon S3 without setting up a server or cluster. Which service is most suitable?

A. Amazon RDS for MySQL

B. Amazon EMR with Spark SQL

C. AWS Glue

D. Amazon Athena

16. Which AWS service can help schedule and orchestrate complex multi-step data pipelines, including dependencies, retries, and parallelization?

A. AWS DataSync

B. AWS Step Functions

C. Amazon Kinesis Data Firehose

D. Amazon Redshift Spectrum

17. A machine learning pipeline needs to be triggered every time new files appear in an S3 bucket. Which AWS feature can be used to invoke an ETL job in response to the S3 event?

A. S3 Event Notifications + AWS Lambda

B. S3 Inventory + AWS Glue

C. S3 Batch Operations + AWS Glue

D. S3 Replication + AWS Step Functions

18. For streaming ingestion with Amazon Kinesis, what is the main difference between Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose?

A. Kinesis Data Firehose supports real-time processing while Kinesis Data Streams only supports batch

B. Kinesis Data Streams is fully managed, while Kinesis Data Firehose is not

C. Kinesis Data Streams can be used for custom processing using consumers, whereas Firehose handles delivery directly to destinations

D. Kinesis Data Firehose can only write to Amazon S3, whereas Kinesis Data Streams can write to multiple services

19. You have a JSON dataset that arrives in real-time. You want to perform windowed aggregations and deliver results to an S3 data lake. Which service provides managed Apache Flink to handle these transformations?

A. Amazon Kinesis Data Analytics

B. AWS Batch

C. Amazon Athena

D. Amazon QuickSight

20. When building a data ingestion pipeline for ML, which of the following is a primary consideration when choosing between batch ingestion and streaming ingestion?

A. Network bandwidth limitations

B. Cost of storing data in S3

C. Frequency and latency requirements of data updates

D. Availability of GPU instances

21. In a machine learning workflow, you need to continuously capture logs from a fleet of web servers and load them into Amazon S3 for batch processing. Which AWS service simplifies this pipeline by automatically handling scaling and data delivery?

A. Amazon Kinesis Data Firehose

B. AWS Step Functions

C. AWS Glue Job

D. Amazon Redshift Spectrum

22. A team wants to use Amazon EMR to transform data stored in Amazon S3 using a Spark job. To optimize cost, they decide to use a transient cluster. Which approach helps ensure minimal data loss when the cluster terminates?

A. Store intermediate data on the cluster's HDFS

B. Write intermediate and final output data back to S3

C. Use Amazon EBS volumes for storing intermediate data

D. Use Amazon DynamoDB for storing all Spark shuffle data

23. Which AWS service offers a serverless environment for running Apache Spark jobs without manually managing clusters?

A. AWS Glue

B. Amazon EMR

C. AWS Batch

D. Amazon Athena

24. Your data pipeline must encrypt data at rest while storing it in Amazon S3. Which approach ensures server-side encryption of data written to S3?

A. Use SSE-KMS by specifying the AWS KMS key in the S3 PutObject request

B. Encrypt data manually using OpenSSL before uploading to S3

C. Store data in an unencrypted S3 bucket, then turn on default encryption after upload

D. Use AWS CloudHSM to wrap all objects in custom hardware encryption

25. Which AWS Glue feature allows you to generate ETL scripts automatically based on your source schema?

A. Glue Triggers

B. Glue Crawlers

C. Glue Development Endpoints

D. Glue DataBrew

26. A company wants to reduce data storage costs by compressing data as part of its batch ETL pipeline in AWS Glue. Which compression format is typically recommended for columnar storage and efficient queries in services like Amazon Athena?

A. Gzip

B. ZIP

C. Parquet with Snappy compression

D. Tar with LZMA compression

27. Which of the following statements about AWS Batch is true?

A. It provides real-time stream processing capabilities

B. It automates the scheduling and execution of containerized batch jobs

C. It is designed primarily for interactive SQL queries

D. It is only used for short-lived jobs and cannot handle large workloads

28. You have a streaming application that needs custom real-time processing logic and low-latency transformations. Which service can directly consume data from Amazon Kinesis Data Streams and run custom Apache Flink applications?

A. Amazon Kinesis Data Firehose

B. AWS Glue Data Catalog

C. Amazon Kinesis Data Analytics (Managed Service for Apache Flink)

D. Amazon EMR

29. When deciding on a storage layer for a high-throughput ML training pipeline, which of the following is most critical to consider?

A. Support for ephemeral instance stores only

B. Integration with Amazon SageMaker built-in algorithms

C. Input/output throughput and parallelism requirements

D. Access to a MySQL engine for real-time queries

30. A large data processing job runs daily and is known to be highly variable in duration. Which AWS service or feature helps right-size compute resources automatically for cost efficiency?

A. AWS Lambda with provisioned concurrency

B. AWS Glue Auto Scaling

C. EMR with Auto Scaling enabled

D. Amazon QuickSight SPICE engine

31. Which statement is most accurate regarding Amazon S3's eventual consistency model?

A. S3 is strongly consistent for all operations by default

B. S3 provides read-after-write consistency for PUTS of new objects, but eventual consistency for overwrite PUTS and DELETES

C. S3 never guarantees read-after-write consistency

D. S3 uses immediate consistency for all operations except HEAD requests

32. A team wants to store a massive volume of data (petabytes) for a data lake, but also needs to query subsets of this data using SQL without loading it into a separate database. Which AWS service supports this requirement?

A. Amazon EMR with Spark SQL

B. AWS Glue ETL Jobs

C. Amazon Athena

D. AWS Batch

33. To enable high-performance parallel reads for ML training, which approach is recommended for data stored in Amazon S3?

A. Use a single large file in S3 to reduce overhead

B. Partition and compress the data into multiple files

C. Store the data in an S3 bucket in a single partition

D. Use plain text CSV files with no compression

34. Your ML pipeline has to ingest data from relational databases, NoSQL stores, and file systems. Which AWS service can catalog all these data sources for consistent schema management?

A. AWS Glue Data Catalog

B. Amazon DynamoDB Streams

C. Amazon RDS Data API

D. Amazon QuickSight

35. You need a fully managed, serverless solution to run SQL queries on streaming data for real-time dashboards. Which service is most appropriate?

A. Amazon Kinesis Data Analytics (SQL Application)

B. Amazon EMR with Hive

C. AWS Glue ETL Jobs

D. Amazon Redshift

36. A data engineer needs to automate the creation of multiple Amazon EMR clusters on a schedule to process different datasets. Which AWS service can help manage the orchestration of these steps, including error handling?

A. AWS Step Functions

B. Amazon Athena

C. Amazon EC2 Auto Scaling

D. AWS Glue Data Catalog

37. Which statement about Amazon EBS is true in the context of data engineering for ML?

A. EBS volumes are automatically replicated across multiple Regions

B. EBS provides a block-level storage volume for use with EC2 instances

C. EBS volumes can be mounted on multiple EC2 instances simultaneously in read-write mode

D. EBS is primarily used for serverless, event-driven data processing

38. You are setting up a data ingestion pipeline for ML training that requires exactly-once processing semantics. Which combination of services or features can help achieve this goal?

A. Amazon Kinesis Data Streams with at-least-once delivery and deduplication logic in the consumer

B. Amazon Managed Streaming for Apache Kafka (MSK) with idempotent producers and transactional consumers

C. Amazon EMR with a custom rollback mechanism

D. AWS Batch with multiple retries

39. A team wants to migrate their nightly ETL job from an on-premises Hadoop cluster to AWS without having to manage servers. Which service offers a managed Hadoop ecosystem with minimal administration overhead?

A. Amazon EMR

B. AWS Glue for Spark

C. Amazon Redshift

D. AWS Batch

40. When designing a data lake on AWS for ML workloads, which approach ensures efficient queries and minimal scanning of irrelevant data?

A. Storing all data in a single bucket with no partitioning

B. Using partitioning strategies based on common query patterns

C. Relying solely on row-level compression formats like Gzip

D. Converting data to CSV for maximum compatibility