1. What does the AWS Glue Metadata Catalog service do?

1 / 1 point

● The AWS Glue Metadata Catalog provides a repository where a company can store, find, and access metadata, and use that metadata to query and transform the data.

○ The AWS Glue Metadata Catalog provides a repository where a company can store and find metadata to keep track of user permissions to data in a data lake.

○ The AWS Glue Metadata Catalog provides a data transformation service where a company can author and run scripts to transform data between data sources and targets.

○ The AWS Glue Metadata Catalog is a query service that uses standard Structured Query Language (SQL) to retrieve data.

✓ **Correct**

AWS Glue Metadata Catalog is the central metadata repository, and it consists of highly scalable collection of tables that are organized into databases. For more information, see the *AWS Glue Data Catalog* video.

2. A solutions architect is working for a customer who wants to build a data lake on AWS to store different types of raw data. Which AWS service should the solutions architect recommend to the customer to meet their requirements?

1 / 1 point

○ AWS Glue Metadata Catalog

○ Amazon OpenSearch Service

○ Amazon EMR

● Amazon Simple Storage Service (Amazon S3)

✓ **Correct**

Amazon S3 stores data contents of any type together in buckets with virtually unlimited storage. This storage type is best suited for data lakes. For more information, see the video Amazon S3. For more information, see the *Amazon S3* video in week 2.

3. Which statement BEST describes batch data ingestion?

1 / 1 point

○ Batch data ingestion is the process of capturing gigabytes (GB) of data per second from multiple sources, such as website clickstreams, database event streams, financial transactions, social media feeds, IT logs, and location-tracking events.

○ Batch data ingestion is a serverless data integration service that makes it easier to discover, prepare, and combine data for analytics, machine learning, and application development.

○ By using batch data ingestion, a user can create a unified metadata repository across various services on AWS.

● Batch data ingestion is the process of collecting and transferring large amounts of data that have already been produced and stored on premises or in the cloud.

✓ **Correct**

Batch-based data ingestion processes large amounts of data that have already been produced or are being ingested periodically. Batch ingestion works best for environments where producing data insights are not time-sensitive. For more information, see the *Batch Data Ingestion with AWS Transfer Family* video.

4. Which service is commonly used for real-time data processing when Amazon Kinesis Data Streams is used for data ingestion?

1 / 1 point

○ Amazon Athena

● Amazon Kinesis Data Analytics

○ Amazon EMR

○ AWS Glue job

✓ **Correct**

Amazon Kinesis Data Analytics processes data streams and generates real-time dashboards. For more information, see the *AWS Services for Analytics* video.

5. Apache Hadoop is an open-source framework that is used to efficiently store and process large datasets. A solutions architect is working for a company that currently uses Apache Hadoop on-premises for data processing jobs. The company wants to use AWS for these jobs, but they also want to continue using the same technology. Which service should the solutions architect choose for this use case?

1 / 1 point

○ Amazon Kinesis Data Analytics

○ Amazon OpenSearch Service

● Amazon EMR

○ AWS Lambda

✓ **Correct**
Amazon EMR is a managed cluster platform that simplifies running big data frameworks—such as Apache Hadoop and Apache Spark—on AWS to process and analyze vast amounts of data. For more information, see the *AWS Services for Data Processing* video in week 2.

6. A team of machine learning (ML) experts are working for a company. The company wants to use the data in their data lake to train an ML model that they create. The company wants the most control that they can have over this model and the environment that it is trained in. Which AWS ML approach should the team take?

1 / 1 point

○ Create an AWS Lambda function with the training logic in the handler, and run the training based on an event.

○ Launch an Amazon Elastic Compute Cloud (Amazon EC2) instance and run Amazon SageMaker on it to train the model.

○ Use a pretrained model from an AWS service, such as Amazon Rekognition.

● Launch an Amazon Elastic Compute Cloud (Amazon EC2) instance by using an AWS Deep Learning Amazon Machine Image (AMI) to host the application that will train the model.

✓ **Correct**
The team of ML experts will probably use EC2 instances for their compute power on AWS. They can launch an EC2 instance with the AWS Deep Learning AMIs that provide the greatest control over

building and managing deep learning models and clusters. For more information, see the *AWS Services for Predictive Analytics and Machine Learning* video in week 1.

7. What is the main value proposition of data lakes?

1 / 1 point

○ The ability to define the data schema before ingesting and storing data.

● The ability to ingest and store data that could be the answer for future questions when they are processed with the correct data processing mechanisms.

○ The ability to combine multiple databases together to expand their capacity and availability.

○ The ability to store user-generated data, such as data from antennas and sensors.

Correct
A data lake is a centralized repository that stores data as-is, without needing to first structure the data and run different types of analytics. The ingested data can be later processed and visualized for specific needs. For more information, see the *Why Data Lakes* video in week 1.

8. Which statements about data lakes and data warehouses are true? (Choose TWO.)

1 / 1 point

□ Data lakes use schema-on-write architectures and data warehouses use schema-on-read architectures.

☑ Data lakes offer more choices in terms of the technology that is used for processing data. In contrast, data warehouses are more restricted to using Structured Query Language (SQL) as the query technology.

Correct
Data lakes provide more power and flexibility by supporting multiple choices for processing data. For more information, see the *Comparison of a Data Lake to a Data Warehouse* video in week 1.

☑ The solutions architect can combine both data lakes and data warehouses to better extract insights and turn data into information.

> **Correct**
> Some common architectures use data lakes to ingest, store, and clean data. Then, the solutions architect can move that data into a data warehouse for visualization. For more information, see the *Comparison of a Data Lake to a Data Warehouse* video in week 1.

□ The solutions architect cannot attach data visualization tools to data warehouses.

□ Data lakes are not future-proof, which means that they must be reconfigured each time new data is ingested.

9. A company plans to explore data lakes and their components. What are reasons to invest in a data lake? (Choose TWO.)

1 / 1 point

□ Increase operational overhead

☑ Offload capacity from databases and data warehouses

> **Correct**
> Databases usually have their storage and processing mechanisms tied together, which makes them less flexible to scale. Therefore, one of the reasons to invest in a data lake is to offload databases and data warehouses to improve performance and save costs. For more information, see the *Why Data Lakes* video in week 1.

□ Limit data movement

□ Make data available from integrated departments

☑ Lower transactional costs

> **Correct**
> One of the most significant advantages of a data lake is being able to store data without needing to think about its structure. This aspect of data lakes provides companies with more cost-effective options to store and scan data. For more information, see the *Why Data Lakes* video in week 1.

10. Which term indicates that a data lake lacks curation, management, cataloging, lifecycle or retention policies, and metadata?

1 / 1 point

● Data swamp

○ Data warehouse

○ Data catalog

○ Database

✓ **Correct**

Data swamp is an informal term that represents a data lake with disorganized data. For more information, see the *Data Lakes Components* video in week 1.

11. Which statement about whether data lakes make it easier to follow the "right tool for the job" approach is TRUE?

1 / 1 point

○ No, data lakes do not make it easier to follow "the right tool for the job approach" because you are tied to a specific AWS service.

● Yes, data lakes make it easier to follow "the right tool for the job" approach because storage can be decoupled from processing and ingestion.

○ No, data lakes do not make it easier to follow "the right tool for the job approach" because data lakes can only handle structured data.

○ Yes, data lakes make it easier to follow "the right tool for the job" approach because data lakes can only handle structured data.

✓ **Correct**

In traditional data warehouse solutions, storage and compute are tightly coupled, which can make it difficult to optimize costs and data processing workflows. With Amazon Simple Storage Service (Amazon S3), users can cost-effectively store all data types in their native formats. Users can then launch as many (or as few) virtual servers as they need by using Amazon Elastic Compute Cloud (Amazon EC2) to run analytical tools. They can also use services in the AWS analytics portfolio—such as Amazon Athena, AWS Lambda, Amazon EMR, and Amazon QuickSight—to process data. For more information, see the *Use the Right Tool for the Job* video.

12. Which scenario represents AWS Glue jobs as the BEST tool for the job?

1 / 1 point

○ Analyze data in real time as data comes into the data lake.

○ Transform data in real time as data comes into the data lake.

○ Analyze data in batches on schedule or on demand.

● Transform data on a schedule or on demand.

✓ **Correct**

An AWS Glue job runs extract, transform, and load (ETL) scripts that connect to your source data, process it, and then write it out to your data target. AWS Glue triggers can start jobs based on a schedule or event, or on demand. For more information, see the *Columnar Data Formats and Amazon Athena Optimizations* reading in week 4.

13. Which task is performed by an AWS Glue crawler?

1 / 1 point

○ Map data from one schema to another schema.

○ Analyze all data in the data lake to create an Apache Hive metastore.

○ Store metadata in a catalog for indexing.

● Populate the AWS Glue Data Catalog with tables.

✓ **Correct**

A crawler can populate the AWS Glue Data Catalog with tables. The crawler is the primary method used by most AWS Glue users. For more information, see the *Using S3, Glue and Athena to Get Insights about NYC Taxi Data* video in week 4.

14. A software developer recently uploaded data logs from their application to Amazon Simple Storage Service (Amazon S3). Who is responsible for encrypting both the data at rest in the S3 bucket and the data in transit to the S3 bucket, according to the AWS shared responsibility model?

1 / 1 point

○ AWS

● Customer

○ Both AWS and the customer

○ Third-party security company

✓ **Correct**
According to the AWS shared responsibility model, customers are responsible for the security in the cloud. AWS provides many features that customers can use to encrypt data both at rest and in transit. For more information about the correct answer, see the *Security and Compliance* reading.

15. What makes Amazon QuickSight different, compared to other traditional business intelligence (BI) tools?

1 / 1 point

○ The ability to create sharable dashboards

● Super-fast, Parallel, In-memory Calculation Engine (SPICE)

○ Data encryption at every layer

○ The ability to visualize data

✓ **Correct**
SPICE is a QuickSight feature that is engineered to rapidly perform advanced calculations and serve data. For more information, see the *Introduction to Amazon QuickSight* video in week 4.

16. What is the purpose of the Registry of Open Data on AWS?

1 / 1 point

○ Provide a service that people can use to ingest software as a service (SaaS) application data into a data lake.

○ Help people discover and share datasets that are available outside of AWS resources.

○ Provide a service that people can use to transform public datasets that are published by data providers through an API.

● Help people discover and share datasets that are available through AWS resources.

✓ **Correct**

The Registry of Open Data on AWS is a collection of publicly available datasets that users can access from AWS resources. For more information, see the *Registry of Open Data* on AWS video in week 4.

17. Which statements about data organization and categorization in data lakes are TRUE? (Choose TWO.)

1 / 1 point

☐ Data lakes are not future-proof, which means that they must be reconfigured each time new data is ingested.

☑ When cataloging data, it is a best practice to organize the data according to the access pattern of the user who will access it.

✓ **Correct**
It is important to find the most appropriate method of categorizing data according to the access pattern. For more information, see the *Use the Right Tool for the Job* video.

☑ Amazon Simple Storage Service (Amazon S3) is mostly used for storage, and AWS Glue is mostly used for categorizing data.

✓ **Correct**
Amazon S3 is an object storage service that stores data as objects within buckets. In contrast, AWS Glue is a fully managed extract, transform, and load (ETL) service that helps users analyze and categorize data. For more information, see the *Use the Right Tool for the Job* video.

☐ Data lakes need to be schema-on-write. In this case, users need to transform all the data before they load it into the data lake.

☐ Users must delete the original raw data to keep their data lake organized and cataloged.

18. Which type of data has the HIGHEST probability of containing structured data?

1 / 1 point

○ Video files from mobile phone photo libraries

○ Raw data from marketing research surveys

○ Customer reviews on products in retailer websites

● Data that is sitting in a relational MySQL table

✓ **Correct**
Structural data is data that is easy for computer systems to consume in its original format, without further modification. A relational database table has schemas, primary keys, foreign keys, and associated data relationships—which means that it works well for storing highly structured data. The database engine is a relational table, and might even reject data that does not fit a rigid structure. For more information, see the *Understanding Data Structure and When to Process Data* video.

19. What is the most common way of categorizing data in terms of structure? **1 / 1 point**

● Structured data, unstructured data, and semi-structured data

○ Ready data, not-ready data, and semi-ready data

○ Development data, quality assurance (QA) data, and production data

○ The good data, the bad data, and the ugly data

✓ **Correct**
Data that is categorized in structured and semi-structured formats have some consistency that makes it easier for computer systems to consume without further modification. In contrast, unstructured data contains content that does not have a predefined data model. For more information, see the *Understanding Data Structure and When to Process Data* video in week 3.

20. Which statement about data consumption in Amazon Kinesis Data Streams is TRUE? **0 / 1 point**

● If data is not consumed within 15 minutes, Kinesis will delete the data that was added to the stream. This case is true even though the data-retention window is greater than 15 minutes.

○ If data is consumed by a consumer, that consumer can never get that same data again. This case is true even if the data is still in the stream, according to the data-retention window.

○ Data consumers must use an AWS SDK to correctly fetch data from Kinesis in the same order that it was ingested. However, AWS Lambda functions do not need to fetch data from Kinesis in a specific order because Lambda integrates natively with AWS services, including Kinesis.

○ Data is automatically pushed to each consumer that is connected to Kinesis. Thus, consumers are notified that new data is available, even when they are not running the Kinesis SDK for data consumption.

⊗ **Incorrect**

By default, the data-retention period in Amazon Kinesis is 24 hours. Users can issue service API calls during the 24-hour data-retention period. For more information about the correct answer, see the *Data Streaming Ingestion with Kinesis Services* video.