



So that teams can spend time with their families, there will be limited Support from Dec 23rd through Jan 1st. Happy Holidays!

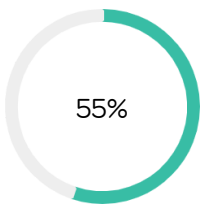
AWS Big Data Specialty Certification - Practice Exam

4 hours65 Questions3.69 Minutes per Question

[Advanced \(/search?type=Practice Exam Challenge&difficulty=Advanced&categories=AWS\)](#)

Go BackStart Challenge

Question ListShow All Answers





Go Back  
**Great Start!**  
You did not pass this challenge on this attempt.

Expectations Report Card

AWS Big Data - Domain 1 - Collection	45.45%
AWS Big Data - Domain 2 - Storage	54.55%
AWS Big Data - Domain 3 - Processing	54.55%
AWS Big Data - Domain 4 - Analysis	63.64%
AWS Big Data - Domain 5 - Visualization	45.45%
AWS Big Data - Domain 6 - Security	70%

Exam Breakdown

AWS Big Data - Domain 1 - Collection

1. Your company releases new features with high frequency while demanding high application availability. As part of the application's A/B testing, logs from each updated Amazon EC2 instance need to be analyzed in near real-time to ensure that the application is working flawlessly after each deployment. If the logs show any abnormal behavior, then the application version of the instance is changed to a more stable one. Which of the following methods should you use for shipping and analyzing the logs in a highly-available manner?  

- A Ship the logs to a large Amazon EC2 instance and analyze the logs in a live manner using AWS Glue.
- B Ship the logs to Amazon S3 for durability and use Amazon EMR to analyze the logs in a batch manner each hour.
- C Ship the logs to an Amazon Kinesis stream and have the consumers analyze the logs in a live manner.**
- D Ship the logs to Amazon CloudWatch Logs and use Amazon EMR to analyze the logs in a batch manner each hour.

**Correct Answer: C****Why is this correct?**

Correct. Kinesis is the best solution to stream logs to analyze and process them in a rapid and highly-available way.

INCORRECT

2. Does AWS Direct Connect allow you access to all Availability Zones within a region?  

- A No**
- B Sometimes, depending on the region.
- C Yes**
- D There are only two Availability Zones per region.

**Your Answer: A****Why is this incorrect?**

Incorrect. Each AWS Direct Connect location enables connectivity to all Availability Zones within the geographically nearest AWS region.

Further Reading: <https://aws.amazon.com/directconnect/faqs/> (<https://aws.amazon.com/directconnect/faqs/>)

**Correct Answer: C****Why is this correct?**

Correct. Each AWS Direct Connect location enables connectivity to all Availability Zones within the geographically nearest AWS region.

Further Reading: <https://aws.amazon.com/directconnect/faqs/> (<https://aws.amazon.com/directconnect/faqs/>)

INCORRECT

3. You need a secure, dedicated connection from your data center to AWS so you can use additional compute resources (EC2) without using the public internet. Which is your best option?



**A** An Amazon Dedicated Connection.

B An encrypted tunnel to VPC

**C** Direct Connect

D None of the above; AWS requires you to connect over the public internet.

**Your Answer: A**

**Why is this incorrect?**

Incorrect. This isn't a real service or product provided by AWS at this time.

**Correct Answer: C**

**Why is this correct?**

Correct. Direct Connect does exactly what is required by the scenario! It allows you to create dedicated connections from your data center to AWS using APNs (usually ISPs in the area that will provide direct connections from your data center to an AWS data center).

4. You need to filter and transform incoming messages coming from a smart sensor you have connected with AWS. Once messages are received, you need to store them as time series data in DynamoDB. Which AWS service can you use?



A IoT Device Shadow Service

B Redshift

C Kinesis

**D** IoT Rules Engine

**Correct Answer: D**

**Why is this correct?**

The IoT rules engine will allow you to send sensor data over to AWS services like DynamoDB

INCORRECT

5. You currently have databases running on-site and in another data center off-site. What service allows you to consolidate to one database in Amazon?



A AWS Kinesis

**B AWS Database Migration Service**

**C AWS Data Pipeline**

D AWS RDS Aurora

### Your Answer: C

#### Why is this incorrect?

Incorrect. Data Pipeline is a useful service when you need to migrate bits of data between sources inside of AWS, but it isn't suited to migrating an entire database from your data center to AWS.

### Correct Answer: B

#### Why is this correct?

Correct. AWS Database Migration Service can migrate your data to and from most of the widely used commercial and open source databases. It supports homogeneous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle to Amazon Aurora. Migrations can be from on-premises databases to Amazon RDS or Amazon EC2, databases running on EC2 to RDS, or vice versa, as well as from one RDS database to another RDS database.

Further Reading: <https://aws.amazon.com/dms/> (<https://aws.amazon.com/dms/>)

6. You have configured an application that batches up data on the servers before submitting it for intake. Your front-end or application server failed, and now you have lost log data. How can you prevent this from occurring in the future while still ensuring that you will have rapid access to your data from multiple different applications?



A Input historical log information using Amazon Machine Learning and use Redshift to analyze and store the logs.

B Trigger Lambda to invoke a process when a log file has been uploaded to Amazon S3 or modified.

**C Submit system and application logs directly to Amazon Kinesis Streams using the Kinesis agent on the front-end and application machines themselves.**

D Submit system and application logs to Amazon EMR and access the data for processing within seconds.

### Correct Answer: C

#### Why is this correct?

Correct. With Amazon Kinesis Streams, you can have producers push data directly into an Amazon Kinesis stream. For example, you can submit system and application logs to Amazon Kinesis Streams and access the stream for processing within seconds. This prevents the log data from being lost if the front-end or application server fails, and reduces local log storage on the source. Amazon Kinesis Streams provides accelerated data intake because you are not batching up the data on the servers before you submit it for intake.

Further Reading: [https://d0.awsstatic.com/whitepapers/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf)  
([https://d0.awsstatic.com/whitepapers/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf))

INCORRECT

7. You are migrating several applications to the cloud from an on-premises environment. You've been asked to select an instance family to use for a relatively well-used microservice as well as the appropriate instance family for a machine learning application. Which instance families do you suggest?



A I for microservices and R for machine learning.

B P instances for the machine learning needs and T instances for the microservice.

C M for machine learning and I for the microservice.

D I instances for machine learning and R for microservices.

### Your Answer: D

#### Why is this incorrect?

Incorrect. R instances are useful for high-performance databases, or Hadoop or Spark clusters, but aren't necessarily the best fit for the GPU-intensive requirements of machine learning applications. I-family instances are better for NoSQL databases.

### Correct Answer: B

#### Why is this correct?

Correct. P instances are well-suited to general-purpose machine learning applications because of their ability to handle general purpose GPU compute loads. T family machines are well suited to general purpose requirements such as microservices.

INCORRECT

8. What combination of services do you need for the following requirements: accelerate petabyte-scale data transfers, load streaming data, and the ability to create scalable, private connections. Select the correct answer order.



A Snowball, Kinesis Firehose, Direct Connect

B Data Migration Services, Kinesis Firehose, Direct Connect

C Snowball, Data Migration Services, Direct Connect

D Snowball, Direct Connection, Kinesis Firehose

### Your Answer: B

#### Why is this incorrect?

Incorrect. Data Migration Services is not a service you would use here.

[https://d0.awsstatic.com/whitepapers/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf)

([https://d0.awsstatic.com/whitepapers/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf))



### Correct Answer: A

#### Why is this correct?

Correct. AWS has many options to help get data into the cloud, including secure devices like AWS Import/Export Snowball to accelerate petabyte-scale data transfers, Amazon Kinesis Firehose to load streaming data, and scalable private connections through AWS Direct Connect.

[https://d0.awsstatic.com/whitepapers/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf)  
([https://d0.awsstatic.com/whitepapers/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf))

INCORRECT

9. Your application requires real-time streaming of data. Each record is 500 KB. It is crucial that the data is delivered and processed as it comes in record-by-record with minimal delay. Which solution allows you to do that?  

A SQS

B RDS

C Spark Streaming

D Kinesis Stream

**Your Answer: A**



**Why is this incorrect?**

Incorrect. SQS cannot handle the payload size.

**Correct Answer: D**

**Why is this correct?**

Correct. SQS cannot handle the payload size, Spark Streaming does it in batches, and RDS cannot do this work.

10. You work for a start-up that tracks commercial delivery trucks via GPS. You receive coordinates that are transmitted from each delivery truck once every 6 seconds. You need to process these coordinates in near real-time from multiple sources and load them into Elasticsearch without significant technical overhead to maintain. Which tool should you use to digest the data?  

A Amazon SQS

B Amazon EMR

C AWS Data Pipeline

D Amazon Kinesis Firehose

**Correct Answer: D**

**Why is this correct?**

Correct. Amazon Kinesis Firehose is the easiest way to load streaming data into AWS. It can capture, transform, and load streaming data into Amazon S3, Amazon Redshift, and Amazon Elasticsearch Service, enabling near real-time analytics with existing business intelligence tools and dashboards.

Further Reading: <https://aws.amazon.com/kinesis/firehose/faqs/> (<https://aws.amazon.com/kinesis/firehose/faqs/>)

11. Your client has a web app that emits multiple events to Amazon Kinesis Streams for reporting purposes. Critical events need to be immediately captured before processing can continue, but informational events do not need to delay processing. What solution should your client use to record these types of events without unnecessarily slowing the application?



- A Log all events using the Kinesis Producer Library.
- B Log critical events using the Kinesis Producer Library, and log informational events using the PutRecords API method.
- C Log critical events using the PutRecords API method, and log informational events using the Kinesis Producer Library.**
- D Log all events using the PutRecords API method.

### Correct Answer: C

#### Why is this correct?

The PutRecords API can be used in code to be synchronous; it will wait for the API request to complete before the application continues. This means you can use it when you need to wait for the critical events to finish logging before continuing. The Kinesis Producer Library is asynchronous and can send many messages without needing to slow down your application. This makes the KPL ideal for the sending of many non-critical alerts asynchronously.

## AWS Big Data - Domain 2 - Storage



INCORRECT

12. A utility company is building an application that stores data coming from more than 10,000 sensors. Each sensor has a unique ID and will send a datapoint (approximately 1 KB) every 10 minutes throughout the day. Each datapoint contains the information coming from the sensor, as well as a timestamp. This company would like to rapidly query information coming from a particular sensor for the past week and delete all of the data that is older than 4 weeks. Using Amazon DynamoDB for its scalability and rapidity, what is the most cost-effective way to implement this?



- A Use one table for each week with a partition key that is the connector between the sensor ID and timestamp.
- B Use one table for each week with a partition key that is the sensor ID and a key for the timestamp.**
- C Use one table with a partition key that is the sensor ID and a sort key that is the timestamp.**
- D Use one table with a partition key that is the concatenation of the sensor ID and timestamp.

### Your Answer: C

**Why is this incorrect?**

Incorrect. Using a single table reduces the ability to query the past week rapidly and quickly deletes older data.

**Correct Answer: B****Why is this correct?**

Correct. This would allow the look up sensor data within a particular time range.

INCORRECT

13. Your DynamoDB items are 1.5 KB in size and you want to write 20 items per second. How many WCUs do you need?



A 40

B 80

C 10

D 20

**Your Answer: D****Why is this incorrect?**

Incorrect. Remember that Writes Capacity Units support one 1 KB-sized item write per second. You also need to round up item sizes to the nearest 1 KB.

**Correct Answer: A****Why is this correct?**

Correct. When doing the calculation, here's what you would do:

$1.5 \text{ KB} / 1 = 1.5 \text{ KB}$ , which gets rounded up to 2.  $2 \text{ RCUs} * 20 \text{ items} = 40 \text{ WCUs}$

Remember that Writes Capacity Units support one 1 KB-sized item write per second. You also need to round up item sizes to the nearest 1 KB.

14. What is true about a Global Secondary Index (GSI) on DynamoDB?



A Only the partition key can be different from the table.

B Only the sort key can be different from the table.

C The partition key and sort key can be different from the table.

D Either the partition key or the sort key can be different from the table, but not both.

**Correct Answer: C****Why is this correct?**



Correct. For GSIs, both the partition and sort key can be different from that of the table. For LSIs, only the sort key can be different.

15. You have a 500-GB file in Amazon S3. Each night, you run a `COPY` command into a 10-node Redshift cluster. How could you prepare the data in order to make the `COPY` command more performant?



**A** Split the file into 500 smaller files.

**B** Convert the file format to CSV format.

**C** Split the file into 10 files of equal size.

**D** Compress the file using `gz` compression.

### Correct Answer: A

#### Why is this correct?

Each node of a Redshift cluster has multiple slices that can each load data in parallel, so multiple files per node will always be more efficient than a single file per node.

### Correct Answer: D

#### Why is this correct?

Compressing files will make them more performant when loading them into Redshift with a `COPY` command.

16. Your mobile application uses a DynamoDB backend to log data. The table has 3 GB of data already in it. The primary key/index is on the device ID of the mobile phone. The application also logs the location of the mobile phone. A new marketing campaign requires a quick lookup for all the phones in a particular area. Also, you have checked CloudWatch, and you are using 90% of the provisioned RCUs and WCUs. How do you make sure you can support the new campaign without any downtime?



**A** Create a GSI on location.

**B** Create an LSI on location.

**C** Increase the RCUs.

**D** Increase the WCUs.

### Correct Answer: A

#### Why is this correct?

Correct. Adding a GSI on location will help query by location for the phone numbers. Adding a GSI will also increase the required RCU and WCUs.

### Correct Answer: C

#### Why is this correct?

Correct. Adding a GSI on location will help query by location for the phone numbers. Adding a GSI will also increase the required RCU and WCUs.

**Correct Answer: D****Why is this correct?**

Correct. Adding a GSI on location will help query by location for the phone numbers. Adding a GSI will also increase the required RCU and WCUs.

---

17. Your sales team uploads sales figures daily. You're designing a solution that has durable storage for these sales figure documents that will also protect against accidental deletions of important documents. Which of these solutions could meet these needs?



- A** Store data in an S3 bucket and enable versioning.
- B Store data in two S3 buckets in different AWS regions.
- C Store data in an EBS volume and create snapshots once a week.
- D Store data on EC2 instance storage.

**Correct Answer: A****Why is this correct?**

This would be the best solution because if S3 objects (the sales figures) are deleted, you can restore from a previous version of the object.

---

18. Your application generates logs that need to be stored in a DynamoDB table. The log contains `user_id`, `event_id`, `timestamp` and `status_code`. You expect to get hundreds of events per user, each with a unique `event_id`. The number of users is expected to grow to 300,000 in two months. You will mainly query the table for the `event_id` for a user during a time frame. What would be the best pick for the partition key and the sort key?



- A 'event\_id' as the partition key. No sort key would be needed.
- B** 'user\_id' as the partition key and 'timestamp' as the sort key.
- C 'user\_id' as the partition key and 'event\_id' as the sort key.
- D 'event\_id' as the partition key and 'user\_id' as the sort key.

**Correct Answer: B****Why is this correct?**

Correct. This will make it easier for you to return all records for a user in the time range and you can extract `event_id` from that.

---

INCORRECT

19. Your client has a high-volume DynamoDB table that serves comment information to an internal API. Currently, the table allows you to query with a composite primary key with `postId` as a hash key and `commentId` as a sort key. Application validation ensures that each item has other fields including `timestamp`, `userId`, and `sentimentScore`. The client has several long-running users, and they would like to provide more effective ways of surfacing posts from them from different time frames. How might the client enable this sort of functionality?



A Create a Global Secondary Index with a hash key of `userId` and a sort key of `timestamp`.

B Create a Local Secondary Index with a hash key of `timestamp` and a sort key of `userId`.

C Create a Local Secondary Index with a hash key of `userId` and a sort key of `timestamp`.

D Create a Global Secondary Index with a hash key of `timestamp` and a sort key of `userId`.

**Your Answer: C****Why is this incorrect?**

Local secondary indexes cannot be created after a table is created. Otherwise, this would be a good way of structuring the index!

**Correct Answer: A****Why is this correct?**

Yes! The hash key or partition key can be set to `userId` so that you can query for all the posts of a single user. Additionally, the `timestamp` sort key will allow the application to query the user's comments based on time.

INCORRECT

20. Which DynamoDB index can be created after the table is created?



A Local Secondary Index

B Global Secondary Index

C None of these.

D Primary Hash Index

**Your Answer: A****Why is this incorrect?**

Incorrect. Local secondary indexes must be created at the same time as the DynamoDB table.

**Correct Answer: B****Why is this correct?**

Correct!

## INCORRECT

21. Your client has an application that is seeing large spikes in traffic on weekends. During those spikes, several of the biggest customers of the client periodically report being unable to load their data. The application in question is a `Vue.js` with a frontend hosted on S3, an API Gateway, and has Lambda powered APIs and a DynamoDB data store. When testing the issue, you discover that smaller clients appear to be able to access data fine even during these spikes. What is the most cost-efficient way to resolve the issue?



- A The S3 Bucket is getting overwhelmed with too many GET requests from the same location. You should make sure to notify AWS of the traffic spikes so they can provide additional capacity for you.
- B DynamoDB requires additional read capacity units. Set up auto-scaling and increase capacity every weekend during the spikes.**
- C The S3 Bucket is getting overwhelmed with too many GET requests from the same location. You should make sure to add object prefixes that introduce randomness.
- D Review the DynamoDB partition keys and determine how you can efficiently randomize them. Then, if necessary, increase read capacity units.

**Your Answer: B****Why is this incorrect?**

Incorrect. While this might solve the issue, it isn't the most cost-effective way to do so. Because only a single client sees this issue, it is likely that a table behind the scenes uses some sort of `customer_id` to partition data. Because DynamoDB tables initially distribute capacity equally between partitions this sort of error may occur for larger customers.

**Correct Answer: D****Why is this correct?**

Correct! Because only a single client sees this issue, it is likely that a table behind the scenes uses some sort of `customer_id` to partition data. Because DynamoDB tables initially distribute capacity equally between partitions, this sort of error may occur for larger customers.

22. You have been asked to ensure that all AWS API calls are collected across your company's AWS account and that they are kept around for 90 days for analysis. After that, they must be able to be restored for 3 years. How can you meet these needs in a scalable, cost-effective way?



- A Enable AWS CloudTrail logging across all accounts to a centralized Amazon S3 bucket with versioning enabled. Set a lifecycle policy to move the data to Amazon Glacier daily, and expire the data after 90 days.
- B Enable CloudTrail logging to a centralized S3 bucket, set a lifecycle policy to move the data to Glacier after 90 days, and expire the data after 3 years.**
- C Enable CloudTrail logging to Glacier, and set a lifecycle policy to expire the data after 3 years.

- D Enable CloudTrail logging in all accounts into S3 buckets, and set a lifecycle policy to expire the data in each bucket after 3 years.

**Correct Answer: B****Why is this correct?**

This meets all the requirements and is cost effective by using Glacier.

## AWS Big Data - Domain 3 - Processing

23. Your EMR cluster uses 12 m4.large instances and runs 24 hours per day, but it is only used for processing and reporting during business hours. Which options can you use to reduce the costs?



A Run 12 d2.8xlarge instead without turn-off.

B Use Spot instances for task nodes when needed.

C Use the ReduceMapper distribution of EMR.

D Migrate the data from HDFS to S3 using S3DistCp and turn off the cluster when not in use.

**Correct Answer: B****Why is this correct?**

Correct. Spot instances are cheaper than regular on-demand instances.

**Correct Answer: D****Why is this correct?**

Correct. This would help reduce costs by not paying for idle EMR time.

INCORRECT

24. You need to implement a solution for customer engagement: You need to write queries that join clickstream data from advertising campaign information stored in a DynamoDB table to identify the most effective categories of ads that are displayed on particular websites. Which services would be your best options for this use case? Choose Two



A SQS

B Kinesis

C AWS Glue

D EMR

**Your Answer: C****Why is this incorrect?**

Incorrect. Glue is a great resource to help manage different metadata for data sources. However, it doesn't help us too much with actually processing all the data we want to process.

**Correct Answer: B****Why is this correct?**

Correct. Kinesis is a great option to collect clickstream data. Amazon EMR clusters can read and process Amazon Kinesis streams directly, using familiar tools in the Hadoop ecosystem such as Hive, Pig, MapReduce, the Hadoop Streaming API, and Cascading. You can also join real-time data from Amazon Kinesis with existing data on Amazon S3, Amazon DynamoDB, and HDFS in a running cluster. You can directly load the data from Amazon EMR to Amazon S3 or DynamoDB for post-processing activities.

Further Reading <http://docs.aws.amazon.com/emr/latest/DeveloperGuide/emr-kinesis.html>

(<http://docs.aws.amazon.com/emr/latest/DeveloperGuide/emr-kinesis.html>)

**Correct Answer: D****Why is this correct?**



Correct. Amazon EMR is a great option to process this sort of data. EMR clusters can read and process Amazon Kinesis streams directly, using familiar tools in the Hadoop ecosystem such as Hive, Pig, MapReduce, the Hadoop Streaming API, and Cascading. You can also join real-time data from Amazon Kinesis with existing data on Amazon S3, Amazon DynamoDB, and HDFS in a running cluster. You can directly load the data from Amazon EMR to Amazon S3 or DynamoDB for post-processing activities.

Further Reading <http://docs.aws.amazon.com/emr/latest/DeveloperGuide/emr-kinesis.html>

(<http://docs.aws.amazon.com/emr/latest/DeveloperGuide/emr-kinesis.html>)

---

INCORRECT

25. Your steaming application requires only-once delivery, and out-of-order data is acceptable as long as the data is processed within 5 seconds. Which solution can be used?  

A Kinesis Streams

B Spark Streaming

C SQS standard queues


D None of these

**Your Answer: C****Why is this incorrect?**

Incorrect. SQS standard queues can't guarantee only-once delivery.

**Correct Answer: B****Why is this correct?**

Correct. Spark has micro-batching but can guarantee only-once-delivery if configured correctly.

- 
26. Your application needs to support terabyte-scale processing of data alongside incoming streaming data. Which big data tool can you use?  

A Amazon Data Pipeline

**B** Amazon EMR with Spark

C Amazon Redshift

D Amazon Machine Learning

**Correct Answer: B****Why is this correct?**

Correct. Using Amazon EMR to run Spark's Machine Learning Library (MLlib) is a common tool for such a use case. You can also use Spark Streaming to process the streaming data component.

Further Reading [https://d0.awsstatic.com/whitepapers/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf)  
([https://d0.awsstatic.com/whitepapers/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf))

INCORRECT

**27.** Your company has decided to use the Amazon Machine Learning service to classify social media posts mentioning your company into two categories: posts requiring a response and posts that do not. You have access to a training dataset of 20,000 posts that each contain things like the timestamp, author, and the full text of the post. You are missing the target labels required for training. How can you effectively create valid target label data?



A Using the a priori probability distribution of the two classes, use Monte-Carlo simulation to generate the labels.

**B** Ask the social media handling team to review each post and provide the label.

C Use the sentiment analysis NLP library to determine whether a post requires a response.

**D** Use the Amazon Mechanical Turk web service to publish Human Intelligence Tasks that ask Turk workers to label the posts.**Your Answer: A****Why is this incorrect?**

Training a machine learning model with the output from another machine learning model might increase the likelihood of error. It would be better to pick a solution that has human-validated data.

**Your Answer: C****Why is this incorrect?**

Training a machine learning model with the output from another machine learning model might increase the likelihood of error. It would be better to pick a solution that has human-validated data.

**Correct Answer: B****Why is this correct?**

This would be one great way to make sure the data is labeled correctly without worrying about possible compounding of errors with the machine learning.

**Correct Answer: D****Why is this correct?**

**Why is this correct?**

This would be one great way to make sure the data is labeled correctly without worrying about possible compounding of errors with the machine learning.

INCORRECT

28. You have advertising campaign information stored in a DynamoDB table. You need to write queries that join clickstream data to identify the most effective categories of ads that are displayed on websites. You also need to support data continuing to be streamed into the table. Which Big Data tools should you use?



A QuickSight

B Kinesis Data Streams

C EMR

D Data Pipeline

**Your Answer: D****Why is this incorrect?**

Incorrect. Data Pipeline can perform ETL and ELT on DynamoDB, but in this case, it's not well suited to the task because you can't use it to bring data from Kinesis into DynamoDB.

**Correct Answer: B****Why is this correct?**

Correct. Amazon EMR clusters can read and process Amazon Kinesis streams directly, using familiar tools in the Hadoop ecosystem such as Hive, Pig, MapReduce, the Hadoop Streaming API, and Cascading. You can also join real-time data from Amazon Kinesis with existing data on Amazon S3, Amazon DynamoDB, and HDFS in a running cluster.

**Correct Answer: C****Why is this correct?**

Correct. Since the data is already in DynamoDB and you want to continue to add to it while analyzing it, you'll want a tool that can leverage that effectively. Because you can directly load the data from EMR into DynamoDB, it is a pretty good option.

29. You work for a tech start-up that has developed a bracelet to track health information for hospitalized children. Each bracelet sends data in JSON format every 6 seconds to analyze and then eventually to create a daily report in a portal for doctors. You need to provide a solution for real-time data analytics that is durable, elastic, and parallel. The results should be stored in JSON so that the frontend can get them and present them to the doctors. Which solution should you select?



A EMR to collect the inbound sensor data, analyze the data from EMR with Amazon Kinesis Analytics, and save the results to DynamoDB.

B SQS to collect the inbound sensor data, analyze the data from SQS with a daily scheduled Data Pipeline, and save the results to a Redshift Cluster.



C S3 to collect the inbound sensor data, analyze the data from S3 with Amazon Kinesis, and save the results to a Microsoft SQL Server RDS instance.



D Amazon Kinesis to collect the inbound sensor data, analyze the data with EMR, and output the results to S3 for eventual consumption by the application.

### Correct Answer: D

#### Why is this correct?

Correct. Amazon Kinesis is the best option to collect the inbound sensor data. You can use EMR as a great option for processing the data and preparing it for consumption. Storing the data in S3 can also be an easy way to retrieve it from the application frontend.

INCORRECT

30. You need to store and process data quickly in a cost-effective manner. You can move data easily from its location on disk to wherever you'd like without needing to stream the data. Also, you do not know how much data you will be handling in 6 months, and your processing needs spike intermittently. Specifically, you need to transform the data that comes in by aggregating the different disparate metrics into summary information. Which Big Data tools should you use?  

A DynamoDB and Redshift

B Kinesis Data Streams and DynamoDB

C S3 and Spark on EMR

D S3 and Amazon Machine Learning

### Your Answer: B



#### Why is this incorrect?

Incorrect. KDS would be great to bring in streaming data, but you don't need it in this situation.

### Correct Answer: C

#### Why is this correct?

Correct. S3 is a great foundational data store, and EMR can scale effectively to suit your needs and load data from S3. Spark is also a great solution to do this sort of aggregation.

31. You work for a photo processing start-up and need the ability to change an image from color to grayscale after it has been uploaded to Amazon S3. How can you configure this in AWS without having to deal with persistent infrastructure?  

A Forecast product demand – use Amazon Machine Learning to track color information to predict future changes.

B Log and data feed intake and processing – with Amazon Kinesis Data Streams, you can have producers push changes directly into an Amazon Kinesis Data Stream.

**C** Real-time file processing – you can trigger Lambda to invoke a process where a file has been uploaded to Amazon S3 or modified.

D Real-time file processing – you can trigger EMR to invoke a process where a file has been uploaded to Amazon S3 or modified.

### Correct Answer: C

#### Why is this correct?

Correct! AWS Lambda can do this sort of task efficiently and run whenever they are uploaded.

32. You have to identify potential fraudulent credit card transactions using Amazon Machine Learning. You have been given historical labeled data that you can use to create your model. You will also need to the ability to tune the model you pick. Which model type should you use?



A Categorical

B Cannot be done using Amazon Machine Learning

**C** Binary

D Regression

### Correct Answer: C

#### Why is this correct?

Correct. Binary models, following from the name, help identify solutions to problems that are one thing or one other thing. In this case, transactions must be determined to be either fraudulent or not fraudulent.

33. Your enterprise application requires key-value storage as the database. The data is expected to be about 10 GB the first month and grow to 2 PB over the next two years. There are no other query requirements at this time. What solution would you recommend?



A Hive on HDFS

B RDS

**C** HBase on HDFS

D Hadoop with Spark

### Correct Answer: C

#### Why is this correct?

Correct! This is specifically what HBase is designed for, and HDFS is flexible enough to allow for the size requirements.

## AWS Big Data - Domain 4 - Analysis



INCORRECT

34. Your data warehouse is running on Redshift. You need to ensure that your cluster can be restored in another region in case of a region failure. What actions can you take to ensure that?



A This feature is not yet available in Redshift.

B Enable Cross-Region snapshots in Redshift.

C Use Lambda to create EBS snapshots.

D Create a manual snapshot.

**Your Answer: A**

**Why is this incorrect?**

Cross-Region snapshots in Redshift are an available feature.

**Correct Answer: B**

**Why is this correct?**

Once Cross-Region snapshots is enabled, Amazon Redshift will automatically incrementally backup your cluster to two AWS Regions.

INCORRECT

35. You have 30 GB of data that needs to be loaded into Redshift. Which of the following will speed up the data ingestion? You also want to be sure that the data lives in more than one place inside of AWS anyway.



A Use `S3DistCp`.

B Store the data already sorted in the sortkey order.

C Copy the data to S3 and use `COPY` to move the data into Redshift.

D Compress the data inside of S3 before loading it into Redshift.

**Your Answer: A**

**Why is this incorrect?**

S3 `COPY` is the fastest loading mechanism of data from S3 to Redshift, so this isn't a good option.

**Correct Answer: B**

**Why is this correct?**

.....  
Loading data already in the sortkey order will allow Redshift to save time and not need to sort data.

**Correct Answer: C**

**Why is this correct?**

S3 `COPY` is the fastest loading mechanism, so making sure to put it in S3 will be the most efficient, and you also want to have a backup of the data there anyway.

**Correct Answer: D**

**Why is this correct?**

Compressing data will make loading it into Redshift more efficiently because it reduces bandwidth requirements.

36. You have been tasked to create an enterprise data warehouse. The data warehouse needs to collect data from each of the three channels' various systems and from public records for weather and economic data. Each data source sends data daily for consumption by the data warehouse. Because each data source may be structured differently, an extract, transform, and load (ETL) process is performed to reformat the data into a common structure. Then, analytics can be performed across data from all sources simultaneously. Which tools shall you implement?



A DynamoDB, Data Pipeline, SQS

B S3, EMR, Data Pipeline, Lambda

**C S3, EMR, Redshift, QuickSight**

D RDS, EMR, Data Pipeline, QuickSight

**Correct Answer: C**

**Why is this correct?**

The first step in this process is getting the data from the many different sources onto Amazon S3. Amazon EMR is used to transform and cleanse the data from the source format into the destination and a format. Each transformation job then puts the formatted and cleaned data onto Amazon S3. Amazon Redshift loads, sorts, distributes, and compresses the data into its tables so that analytical queries can execute efficiently and in parallel. For visualizing the analytics, Amazon QuickSight can be used, or one of the many partner visualization platforms via the ODBC/JDBC connection to Amazon Redshift.

37. You have a lot of data (over 50 TB) in your on-premises data warehouse that you need to load into Amazon Redshift. Which option would allow you to load this data in Redshift?



A Configure Data Pipeline to send information over a Direct Connect Connection.

B Use VPC Peering to connect your on-premises network to AWS and send data via that connection.

C Use AWS Glue to create transfer jobs.

**D Snowball into S3 then `COPY` to Redshift.**

**Correct Answer: D**

**Correct Answer: C****Why is this correct?**

You can use AWS Snowball service to transfer the data to Amazon S3 using portable storage devices and then use the `COPY` command to load the data into Redshift effectively.

38. You have a Redshift table that you are designing called 'item\_description' that contains 3MB of data, and you will use it frequently in joins. The table itself isn't frequently updated. What `DISTSTYLE` for the table will optimize queries?



A Change the `DISTSTYLE` to `PARTITION`.

B Change the `DISTSTYLE` to `KEY`.

**C** Change the `DISTSTYLE` to `ALL`.

D Change the `DISTSTYLE` to `EVEN`.

**Correct Answer: C****Why is this correct?**

`DISTSTYLE ALL` will place table data on the first slice of every node in the cluster (this assumes that the table is small enough to do that). It is also useful because the table isn't frequently updated.

39. You need real-time reporting on logs generated from your applications. In addition, you need anomaly detection. The processing latency needs to be one second or less. Which option would you choose if your team has no experience with Machine learning libraries and doesn't want to have to maintain any software installations yourself?



**A** Kinesis Streams with Kinesis Analytics

B Kafka

C Kinesis Firehose to S3 and Athena

D Spark Streaming with SparkSQL and MLlib

**Correct Answer: A****Why is this correct?**

The Kinesis Streams with Kinesis Analytics solution is the one with the lowest latency

**INCORRECT**

40. You need to analyze a large set of JSON data from Kinesis and DynamoDB by querying for a variety of different values inside of the documents to search for particular records. The fields you need to query, and the records themselves, vary significantly. Which Big Data tool should you use if your



organization is trying to use more managed services when possible?

A EMR

B Redshift

C QuickSight

D Elasticsearch

### Your Answer: C

#### Why is this incorrect?

QuickSight is ideal when you already have the data you want to visualize all sorted away. When you're querying for data, you'll want to use a different tool.

### Correct Answer: D

#### Why is this correct?

Amazon ES is ideal for querying and searching large amounts of data. Organizations can use Amazon ES to do the following:

- Analyze activity logs, such as logs for customer-facing applications or websites
- Analyze CloudWatch logs with Elasticsearch
- Analyze product usage data coming from various services and systems
- Analyze social media sentiments and CRM data, and find trends for brands and products
- Analyze data stream updates from other AWS services, such as Amazon Kinesis Streams and DynamoDB
- Provide customers with a rich search and navigation experience
- Monitor usage for mobile applications

INCORRECT

41. A new client is requesting a tool that will provide fast query performance for enterprise reporting and business intelligence workloads, particularly those involving extremely complex SQL with multiple joins and sub-queries. They also want the ability to give analysts access to a central system through tradition SQL clients that allow them to explore and familiarize themselves with the data. What solution do you initially recommend they investigate?



A SQS

B Redshift

C Athena

D EMR

**Your Answer: C****Why is this incorrect?**

While Athena might be an ok fit for this need, the different setup of data sources might make it less familiar to analysts exploring the data.

**Correct Answer: B****Why is this correct?**

Correct! Redshift can fill these needs.

42. Which single action can speed up this query: "SELECT count(\*) FROM transactions WHERE date\_of\_purchase BETWEEN '2017-04-01' AND '2017-05-01' " when it runs on a Redshift table with 10 million rows. The table was created from S3 data with a COPY command.



A Create a sort key on the column `date_of_purchase` .

B Use `date_of_purchase` as the DISTKEY.

C Use LZ0 compression on the `date_of_purchase` column.

D Use `date_of_purchase` as the PARTITION KEY.

**Correct Answer: A****Why is this correct?**

Correct! Sort keys will allow for more efficient queries when you are running filters on that field!

43. Which tool provides the easiest way to run ad-hoc queries for data in S3 without the need to set up or manage any servers.



A SQS

B EMR

C Athena

D Redshift

**Correct Answer: C****Why is this correct?**

Query services like Amazon Athena, data warehouses like Amazon Redshift, and sophisticated data processing frameworks like Amazon EMR all address different needs and use cases; you just need to choose the right tool for the job. Amazon Athena provides the easiest way to run ad-hoc queries for data in S3 without the need to set up or manage any servers.

Further Reading <https://aws.amazon.com/athena/faqs/> (<https://aws.amazon.com/athena/faqs/>)

44. Athena supports which file formats by default?



A JSON

B PDF

C CSV

D Apache Parquet

**Correct Answer: A**

**Why is this correct?**

JSON is supported by Athena.

**Correct Answer: C**

**Why is this correct?**

CSV is a supported format!

**Correct Answer: D**

**Why is this correct?**

Parquet is definitely a supported and recommended format!

## AWS Big Data - Domain 5 - Visualization



INCORRECT

45. Management has requested a comparison of total sales performance in the five North American regions in January. They're hoping to determine how to allocate a budget to regions based on performance in that single period. What sort of visualization do you use in Amazon QuickSight?



A A bar chart

B A line chart

C A stacked area chart

D A histogram

**Your Answer: D**

**Why is this incorrect?**



Histograms are great when you want to compare different non-chronological buckets of data. For example, pre-defined customer segments. We're not doing that here.

### Correct Answer: A

#### Why is this correct?

Bar charts are one of the best visualizations to use to compare multiple types of different data in the same period. A column chart is also a possibility, but that is not listed here.

46. You have a JSON data file in S3 that you are attempting to load into a JavaScript visualization you are writing locally. This visualization makes an HTTP GET request to the S3 location that fails. However, when you attempt to visit the URL being requested by the JavaScript directly from inside your browser, it seems to be loading fine. You are also using a private/incognito window and are not signed into the AWS console. What is the most likely issue?



- A** The CORS settings are preventing the JavaScript from loading the file.
- B The ACLs on the bucket are preventing the JavaScript from loading the file.
- C The bucket policies are preventing the JavaScript from loading the file.
- D The IAM role you used to create and upload the JSON data in the S3 bucket is preventing the JavaScript from loading the file.

### Correct Answer: A

#### Why is this correct?

Yes! This is the most likely issue as the CORS settings may prevent the browser from getting this data

47. You need to visualize data from Spark and Hive running on an EMR cluster. Which of the options is best for an interactive and collaborative notebook for data exploration?



- A Hive
- B D3.js
- C Kinesis Analytics
- D** Zeppelin

### Correct Answer: D

#### Why is this correct?

Apache Zeppelin is an open source GUI which creates interactive and collaborative notebooks for data exploration using Spark. You can use Scala, Python, SQL (using Spark SQL), or HiveQL to manipulate data and quickly visualize results.

INCORRECT

48. You've been asked by the VP of People to showcase the current percentage breakdown of the headcount for each department within your organization. What is the best chart to select to make it easy to compare each department?



A A line chart

B A scatter plot

C A column chart

D A pie chart

**Your Answer: A****Why is this incorrect?**

In this case, there is no chronological data to compare! The line would just be a set of dots!

**Correct Answer: D****Why is this correct?**

In this example, you need to compare a static dataset that makes up a whole. A pie chart is an appropriate chart for this purpose.

49. You need to provide customers with rich visualizations that allow you to easily connect multiple disparate data sources in S3, Redshift, and several CSV files. Which tool should you use that requires the least setup?



A Hue on EMR

B Redshift

C QuickSight

D Elasticsearch

**Correct Answer: C****Why is this correct?**

Amazon Quicksight is the best option to connect the different data sources. While Elasticsearch can use Kibana to visualize data, it would be more work to ingest data into Elasticsearch from those sources, and the same goes for EMR and Hue.

INCORRECT

50. You are attempting to determine if there is any relationship between certain marketing expenditures and the performance of your products. You decide the best way to visualize this is with what kind of



chart?

A Histogram

B Bar Chart

C Scatter Plot

D Stacked Area Chart

**Your Answer: D**

**Why is this incorrect?**



Stacked area charts are great to help compare how data changes over time in relation to other portions of the data.

**Correct Answer: C**

**Why is this correct?**

The scatter plot chart can be useful when you want to plot two different variables and see if a relationship exists between them.

INCORRECT

51. You've been asked to find a solution to visualize company JIRA data alongside GitHub PRs in a way that minimizes developer time. What solution do you propose?  

A Pull JIRA and GitHub data into QuickSight and build visualizations on top of those data sources.

B Pull JIRA and GitHub data into Hue and build the visualizations on top of that.

C Pull JIRA and GitHub data into EMR and build the visualizations after cleaning the data with a transient cluster.

D Pull JIRA and GitHub data into S3 with a few simple Lambda functions, make the data files public and build the visualizations with D3.js.

**Your Answer: D**



**Why is this incorrect?**

This adds a bit of complexity that isn't needed; you can just load data into Quicksight directly!

**Correct Answer: A**

**Why is this correct?**

QuickSight is the only option that has integrations to pull this data in with minimal developer time.

52. Your company recently purchased five different companies that run different backend databases that include Redshift, MySQL, Hive on EMR and PostgreSQL. You need a single tool that can run queries on all the different platform for your daily ad-hoc analysis. Which tools enable you to do that?  

**A** Presto

B QuickSight

C Ganglia

D YARN

### Correct Answer: A

#### Why is this correct?

A single Presto query can process data from multiple sources.

INCORRECT

**53.** You work for a global marketing SaaS vendor that sells to content and marketing managers around the world so they can see analytics about their data. Your frontend development team is attempting to put together automated visualizations for your clients within their dashboards. What solution do you recommend they investigate?



A Kibana

**B** Highcharts

**C** QuickSight

D Hue

### Your Answer: C

#### Why is this incorrect?

QuickSight is more of a solution for organizations themselves to deploy in relation to the data they collect, so it doesn't fit.

### Correct Answer: B

#### Why is this correct?

While these are all one form of visualization tool or another, in this particular case it appears like the frontend development team would require a JavaScript library in order to build a new visualization for the non-technical customers. So, the Highcharts JavaScript library is perfectly suited to this purpose.

**54.** You've been asked by management to bucket customers who are currently in different phases of onboarding. They'd like to see the number of customers in each phase. What sort of visualization type do you use?



A A scatter plot

**B** A histogram

C A stacked 100% area chart

D A bubble chart

### Correct Answer: B

#### Why is this correct?

A histogram is well suited to this because it buckets different sets of data in non-chronological comparisons like this.

INCORRECT

55. You've been asked to select a tool that can easily visualize sales data that comes in as JSON to S3, occasionally as ad-hoc CSV files, and even from the Amazon Redshift data warehouse. The solution must allow multiple users from the finance department to easily access it and occasionally upload their own Excel spreadsheets to compare with existing data. What solution do you recommend?



A Use Kibana and a combination of an S3 bucket that accepts the XLSX downloads and processes them with Lambda to transform them into JSON and index them in Elasticsearch.

B Use Kibana and Amazon Athena to process the S3 data and XLSX files before indexing them in Elasticsearch.

**C** QuickSight and a combination of data source connections with the Redshift cluster and existing S3 JSON documents while still allowing finance to upload files directly.

D Use QuickSight and a combination of data source connections with the Redshift cluster and existing S3 JSON documents along with a Lambda function to process the XLSX files and transform them into a QuickSight-readable format.

### Your Answer: D

#### Why is this incorrect?

The Lambda function here adds a bit of extra work that isn't needed. QuickSight already reads XLSX by default.

### Correct Answer: C

#### Why is this correct?

This solution can easily accomplish all the requirements without the extra work of integrating a bunch of extra tools. Also, QuickSight also supports XLSX files by default!

## AWS Big Data - Domain 6 - Security



INCORRECT

56. Your application development team is building a solution with two applications. The security team wants each application's logs to be captured in two different places because one of the applications produces logs with sensitive data. How can you meet the requirements with the least risk and effort?



- A Aggregate logs into one file, then use Amazon CloudWatch Logs and then design two CloudWatch metric filters to filter sensitive data from the logs.
- B Use Amazon CloudWatch logs to capture all logs, write an AWS Lambda function that parses the log file, and move sensitive data to a different log.**
- C Add logic to the application that saves sensitive data logs on the Amazon EC2 instances' local storage, and write a batch script that logs into the EC2 instances and moves sensitive logs to a secure location.
- D Use Amazon CloudWatch logs with two log groups, one for each application, and use an AWS IAM policy to control access to the log groups as required.**

### Your Answer: B

#### Why is this incorrect?



This would require a bit of extra work and doesn't seem to have an effective way of controlling who accesses what types of logs.

### Correct Answer: D

#### Why is this correct?

This does exactly what is needed and efficiently uses IAM policies to control access to log groups.

INCORRECT

57. Your data analytics team needs to load data into Redshift from S3. Currently, company policy restricts AWS user accounts to developers and not your data engineers. Instead, they each have different Redshift user accounts with access to the cluster. How do you empower them to do their jobs?  

- A You should create an IAM role and attach it to the cluster and make sure it can be used by the specific team members you would like to use it.**
- B Redshift should already have permissions after the cluster is created. Ask them to run a `COPY` command to load in the data.
- C You should create API keys for each Redshift user and have them use those keys to copy data in from S3.
- D You should make the files in the S3 bucket public when reading from a specific IP so that the cluster can access them and load them into Redshift.**

### Your Answer: D

#### Why is this incorrect?

While this *might* work, it's not a very effective solution as it would require a lot of additional maintenance on the S3 buckets rather than the cluster itself. Stick to assigning permissions to the cluster and its users.

### Correct Answer: A

#### Why is this correct?

Redshift's `COPY` commands require some form of authorization to copy the data into a Redshift table from S3. This

solution would allow the **Redshift** users (not IAM users) to have the proper permissions.

58. You have an application with several hundred IoT devices all sending data into S3. Your team has created a mobile application that relies on reading data from DynamoDB. How could you give each mobile device permissions to read that data from DynamoDB?



- A Create an IAM user.
- B Connect to an EC2 instance which will pull the data from DynamoDB securely.
- C Add an encrypted username and password into the app code and decrypt it at runtime.
- D Create an IAM role that can be assumed by an app that allows federated users.**

### Correct Answer: D

#### Why is this correct?

It is bad practice to store any API credentials in a mobile application. Each mobile device should have their own permissions and access credentials to DynamoDB. In order to facilitate this, you can integrate federated users (Facebook, Google, Twitter, Amazon, etc.) credentials with IAM. After authenticated as a federated user, the user/app can then assume an IAM role with the proper read/write permissions to DynamoDB.

59. A mobile application collects data and stores it in multiple Availability Zones within five minutes of being captured in the app. How can you securely meet these requirements?



- A The mobile app should call a REST-based service that stores data on Amazon EBS. Deploy the service on multiple EC2 instances across two Availability Zones.
- B The mobile app should authenticate with an Amazon Cognito identity that is authorized to write to an Amazon Kinesis Firehose with an Amazon S3 destination.**
- C The mobile app should write to an S3 bucket that allows anonymous PutObject calls.
- D The mobile app should authenticate with an embedded IAM access key that is authorized to write to an Amazon Kinesis Firehose with an Amazon S3 destination.

### Correct Answer: B

#### Why is this correct?

Cognito will allow you to securely authenticate pools of users on any type of device.

60. You are collecting large amounts of data from an application that is running on EC2 instances. This application processes sensitive information stored on S3. You access this data over the internet, but your security team is concerned that the internet connectivity to Amazon S3 is a security risk. How could you mitigate this?



A Access the data through a VPN connection.

**B** Access the data through a VPC endpoint for Amazon S3.

C Access the data through an Internet Gateway.

D Access the data through a NAT Gateway.

### Correct Answer: B

#### Why is this correct?

Yes! VPC endpoints for Amazon S3 provide secure connections to S3 buckets that do not require a gateway or NAT instances.

61. What is the result of the following bucket policy?



```
{
  "Statement": [
    {
      "Sid": "Sid2",
      "Action": "s3:*",
      "Effect": "Allow",
      "Resource": "arn:aws:s3:::mybucket/*.",
      "Condition": {
        "ArnEquals": {
          "s3:prefix": "data_team_"
        }
      },
      "Principal": {
        "AWS": [
          "*"
        ]
      }
    }
  ]
}
```

A It will deny all actions if the object prefix is `data_team_`.

B It will allow all actions if the object is in the finance subdirectory of `mybucket`.

**C** It will allow all actions only against objects with the prefix `data_team_` in the `mybucket` bucket.



D It allows access to objects in the `data_team_` bucket namespace.

### Correct Answer: C

#### Why is this correct?



This would pertain to all objects with keys beginning with `data_team_`.

62. You're launching a test Elasticsearch cluster with the Amazon Elasticsearch Service, and you'd like to restrict access to only your office desktop computer that you occasionally share with an intern to allow her to get more experience interacting with Elasticsearch. What's the easiest way to do this?  



- A Create a username and password combination to allow you to sign into the cluster.
- B Create an SSH key and add that to the accepted keys of the Elasticsearch cluster. Then store that SSH key on your desktop and use it to sign in.
- C Create an IAM user and role that allows access to the Elasticsearch cluster.
- D Create an IP-based resource policy on the Elasticsearch cluster that allows access to requests coming from the IP of the machine.**

### Correct Answer: D

#### Why is this correct?

The IP-based policy is part of the cluster creation process and will accomplish what is required.

INCORRECT

63. You have to do a security audit on an EMR cluster running Spark. Which of the following are components of the security of using EMR that you should either be aware of or make sure are enabled?  

- A Hive encryption.**
- B In-transit data encryption between S3 and EMRFS.
- C EC2 instances using encrypted EBS volumes.
- D Server-side encryption on S3 using `S3-SSE/KMS/Custom`.

### Your Answer: A

#### Why is this incorrect?

While Hive does have encryption settings, when running Spark on EMR, it isn't something you need to worry about.

### Correct Answer: B

#### Why is this correct?

Yes, in-transit encryption is another thing that is important to be aware of when working with EMRFS and S3. TLS effectively provides this protection.

### Correct Answer: C

#### Why is this correct?

Yes, the underlying storage of the EC2 instances inside the EMR cluster should be encrypted, and you can do this by encrypting the EBS volumes that store the data.

**Correct Answer: D**

**Why is this correct?**

Server-side encryption is another excellent choice to help ensure the security of your data.

64. You have 30 dedicated Kinesis Stream for uniquely named streaming events. What action can you take so that the Kinesis charges are separated out on the Amazon invoice at the end of the month?



A Enable CloudWatch to monitor the streams.

B Submit a support request to Amazon inside the AWS console.

C Move each named streaming event into a separate AWS account and use consolidated billing.

**D Tag the streams with the streaming event name.**

**Correct Answer: D**

**Why is this correct?**

This does allow you to see the charges split up at the end of the month.

65. Your company needs to design a data warehouse for a client in the retail industry. The data warehouse will store historic purchases in Amazon Redshift. To comply with PCI:DSS requirements and meet data protection standards, the data must be encrypted at rest and have keys managed by a corporate on-premises HSM. How can you meet these requirements in a cost-effective manner?



A Use AWS Import/Export to import a company HSM device into AWS alongside the Amazon Redshift cluster, and configure Redshift to use the imported HSM.

**B Create a VPN connection between a VPC you create in AWS and an on-premises network. Then launch the Redshift cluster in the VPC, and configure it to use your corporate HSM.**

C Use the AWS CloudHSM service to establish a trust relationship between the CloudHSM and the corporate HSM over a Direct Connect connection. Configure Amazon Redshift to use the CloudHSM device.

D Configure the AWS Key Management Service to point to the corporate HSM device, and then launch the Amazon Redshift cluster with the KMS managing the encryption keys.

**Correct Answer: B**

**Why is this correct?**

Redshift can leverage on-premises HSMs for key management using VPN. This meets the requirements by making sure the encryption keys are locally managed.

