90.82%

Congratulations, you passed!

Great job on the challenge, you just earned rewards to use on more challenges.

Add to profile (https://www.linkedin.com/profile/add?startTask=AWS Big Data Specialty Certification - Practice Exam)

Linux Academy

Go Back

Report Card

Expectations	Score
1. AWS Big Data - Domain 1 - Collection	90.91 %
2. AWS Big Data - Domain 2 - Storage	90.91 %
3. AWS Big Data - Domain 3 - Processing	81.82 %
4. AWS Big Data - Domain 4 - Analysis	100 %

5. AWS Big Data - Domain 5 - Visualization	90.91 %
6. AWS Big Data - Domain 6 - Security	90 %

Exam Breakdown

AWS Big Data - Domain 1 - Collection

1. You currently have an on-premises Oracle database and have decided to leverage AWS and use Aurora. You need to do this as quickly as possible. How do you achieve this?

A. It is not possible to migrate an on-premises database to AWS at this time.	
B. Use AWS Data Pipeline to create a target database, migrate the database schema, set up the data replication process, initiate the full load and a subsequent change data capture and apply, and conclude with a switchover of your production environment to the new database once the target database is caught up with the source database.	
C. Use AWS Database Migration Services and create a target database, migrate the database schema, set up the data replication process, initiate the full load and a subsequent change data capture and apply, and conclude with a switch-over of your production environment to the new database once the target database is caught up with the source database.	✓ **Correct**
D. Use AWS Glue to crawl the on-premises database schemas and then migrate them into AWS with Data Pipeline jobs.	

2. 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.	✓ **Correct**
C. M for machine learning and I for the microservice.	
D. I instances for machine learning and R for microservices.	

3. **You need to migrate data to AWS. It is estimated that the data transfer will take over a month via the current AWS Direct Connect connection your company has set up. Which AWS tool should you use?**

A. Establish additional Direct Connect connections.	
B. Use Data Pipeline to migrate the data in bulk to S3.	
C. Use Kinesis Firehose to stream all new and existing data into S3.	
D. Snowball	✓ **Correct**

4. **There is a five-day car rally race across Europe. The race coordinators are using a Kinesis stream and IoT sensors to monitor the movement of the cars. Each car has a sensor and data is getting back to the stream with the default stream settings. On the last day of the rally, data is sent to S3.**

When you go to interpret the data in S3, there is only data for the last day and nothing for the first 4 days. Which of the following is the most probable cause of this?

A. You did not have versioning enabled and would need to create individual buckets to prevent the data from being overwritten.
B. Data records are only accessible for a default of 24 hours from the time they are added to a stream. ✓ **Correct**
C. One of the sensors failed, so there was no data to record.
D. You needed to use EMR to send the data to S3; Kinesis Streams are only compatible with DynamoDB.

5. **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 ✓ **Correct**

6. **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	✓ **Correct**
D. None of the above; AWS requires you to connect over the public internet.	

7. **There is a 14-day backpacking tour across Europe. The tour coordinators are using a Kinesis Data Stream and IoT sensors to monitor the movement of the group. You have changed the default settings on the stream to the max settings. Each backpack has a sensor and data is getting back to the stream with the default stream settings. On the last day of the tour, data is sent to S3. When you go to interpret the data in S3, there is only data for 7 days. Which of the following is the most probable cause of this?**

A. You needed to use EMR to send the data to S3. Kinesis Streams are only compatible with DynamoDB.	
B. One of the sensors failed, so there was no data to record.	
C. You did not have versioning enabled and would need to create individual buckets to prevent the data from being overwritten.	
D. Data records are only accessible up to 7 days from the time they are added to a stream.	✓ **Correct**

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

9. **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	✗ **Your Answer**

Why is this incorrect?

Incorrect. While Kinesis could technically be used as an intermediary between different sources, it isn't a great way to get data into DynamoDB from an IoT device.

D. IoT Rules Engine ✓ **Correct**

Why is this correct?
The IoT rules engine will allow you to send sensor data over to AWS services like DynamoDB

10. You have EC2 instances that you need to connect to your on-premises data center. You need to be able to support a connection speed of 200 Mbps. How should you configure this?

A. Allocate EIPs and an Internet Gateway for your VPC instances, then provision a VPN connection between a VPC and your data center.

B. Provision a VPN connection between a VPC and data center, Submit a Direct Connect partner request to provision cross-connects between your data center and the Direct Connect location, then cut over from the VPN connection to one or more Direct Connect connections as needed. ✓ **Correct**

C. Create an internal ELB for your application, submit a Direct Connect request to provision a 1 Gbps cross-connect between your data center and VPC, then increase the number or size of your Direct Connect connections as needed.

D. Use Direct Connect to provision a 1 Gbps cross-connect between your data center and VPC, then increase the number or size of your Direct Connect connections as needed.

11.

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. ✓ **Correct**
D. Ship the logs to Amazon CloudWatch Logs and use Amazon EMR to analyze the logs in a batch manner each hour.

AWS Big Data - Domain 2 - Storage

1. **Your client's application logs data in large files and runs weekly analytics on these logs for internal reporting for six months after the logs are generated. After six months, the logs are infrequently accessed for up to a year. The client also has a regulatory requirement to store application logs for seven years. How should the company achieve these requirements in the most cost-efficient way?**

A. Store the files in S3 Glacier with a Deny Delete vault lock policy for archives less than seven years old, a vault access policy that restricts read access to the analytics IAM group, and write access to the log writer service role.
B. Store the files in S3 Standard with lifecycle policies to transition the storage class to Standard – IA after six months and delete them after a year. Simultaneously store the files in Amazon Glacier with a Deny Delete ✓ **Correct**

vault lock policy for archives less than seven years old.

C. Store the files in S3 Standard with a lifecycle policy to transition the storage class to Standard - IA after three months. After a year, transition the files to Glacier and add a Deny Delete vault lock policy for archives less than seven years old.

D. Store the files in S3 Standard with a lifecycle policy to remove them after a year. Simultaneously store the files in Amazon S3 Glacier with a Deny Delete vault lock policy for archives less than seven years old.

2. You have created a DynamoDB table with `CustomerID` as the primary key for the table. You need to find all customers that live in a particular ZIP code. How should you configure this?

A. Use `ZipCode` as the partition key for a local secondary index, since there are a lot of ZIP codes and you will probably have a lot of customers. Change `CustomerID` to the global secondary index.

B. Use `ZipCode` as the partition key for a local secondary index, since there are a lot of ZIP codes and you will probably have a lot of customers.

C. Change the primary key to `ZipCode` and use `CustomerID` as the global secondary index.

D. Use `ZipCode` as the partition key for a global secondary index, since there are a lot of ZIP codes and you will probably have a lot of customers. ✓ **Correct**

3. 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.	✓ **Correct**
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.	

4. 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`.	✓ **Correct**
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`.	

5. **An application requires a highly available relational database with an initial storage capacity of 8 TB. The database will grow by 8 GB every day. To support expected traffic, at least eight read replicas will be required to handle database reads. With what service could you meet these requirements?**

A. Amazon Aurora	✓ **Correct**
B. DynamoDB	
C. Amazon S3	
D. Amazon Redshift	

6. **Your company creates mobile games that use DynamoDB as the backend data store to save the high scores. There is a hash and range key on the main table, where the game is the partition key and the username is the sort key. Your highest selling game customers complain that your game slows down to a halt when trying to send the high scores to your backend. CloudWatch metrics suggest you are not exceeding your provisioned WCUs. Your company is currently undergoing an in-depth re-platforming and is wondering how they can improve this situation in the long term. Which option can improve the user experience without increasing costs?**

A. Change the partition key to use just the game.
B. Provision more WCUs.

C. Don't do anything at all.

D. Recreate the table with the username as the partition key and the game as the sort key. ✓ **Correct**

7. **Your company is designing a web application using stateless web servers. Which services would work to store session state data?**

A. Storage Gateway

B. DynamoDB ✓ **Correct**

C. CloudWatch

D. ElastiCache ✓ **Correct**

E. Elastic Load Balancing

8. **You need to design a solution that can return user profile data to your application with millisecond latency or better and that can store this information for two million users. Currently, user profile information is capped at 15 KB and typically doesn't exceed 3 KB. It is very important that this profile information always reflects the most recent changes when read. You also want a way to process changes from user profiles and potentially send out emails when specific types of changes are made but the amount of compute capacity required for this is highly sporadic. What could you use to help build a cost-effective solution to your application?**

A. Use separate S3 objects to store user profile information and run ECS jobs periodically hash the files and compare them to a RDS database of object hashes to determine if they need to be processed and emails might need to be sent out using SES.	
B. Use a DynamoDB table to store profile information as items and then enable DynamoDB streams on the table. Whenever changes are made, evaluate them with a Lambda function and send an email with SES if that is appropriate.	✓ **Correct**
C. Use a DynamoDB table to store profile information as items and then enable DynamoDB streams on the table. Write an application that sits on a cluster of EC2 Reserved Instances and use that to process the DynamoDB streams data and send emails using SES when appropriate.	
D. Use separate S3 objects to store user profile information and use AWS Lambda functions that trigger whenever objects are updated. Then send the customized HTML emails out from SNS to users if appropriate.	

9. **You are using a MapReduce job to analyze the activation of an item you sell. The job is able to tolerate interruptions and occasional instance termination. The number of activations is usually steady throughout the year, except the week before Christmas, where there is a 20X increase. You need to be sure you have both a solution that will consistently provide low-latency performance and one that will allow you to expand processing power significantly when needed to process the additional data. What is the most cost-effective and performance-optimized solution?**

A. Use Amazon DynamoDB and Amazon Elastic MapReduce with Reserved instances.	
B. Amazon RDS and Amazon Elastic MapReduce with Spot instances.	
C. Amazon DynamoDB and Amazon Elastic MapReduce with Spot instances.	✓ **Correct**

D. Amazon RDS and Amazon Elastic MapReduce with Reserved instances.

10. 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. ✓ **Correct**

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.

11. Your client needs to deploy a big data solution so that data analysts can access all company data in a central S3 bucket. Data needs to be segregated by each company department with prefixes, and the analysts must only be able to access the data from their own department. The company is currently using Microsoft Active Directory (AD) and needs to leverage that to manage access to the S3 data. Which solution is right for them?

A. Use the AWS AD Synergization service to create IAM groups based on AD information.

B. Create bucket policies on the S3 bucket that will only allow access to data based on the users' AD ✗ **Your Answer**

group name.

Why is this incorrect?
The S3 bucket policy wouldn't be able to tell the users' AD group name.

C. Leverage the Amazon S3 API with AD credentials to allow users to call directly to the HTTP endpoints to access data.

D. Use IAM Federation functions, and depending on the users' groups in AD, specify an associated role to grant to that user. ✓ **Correct**

Why is this correct?
This would allow access to data to be tied to AD groups, and then the correct role to be applied to the user depending on their AD role.

AWS Big Data - Domain 3 - Processing

1. **You are provisioning an application using EMR. You have requested 100 instances. You are charged $0.015 per hour, per instance. In the first 10 minutes after your launch request, Amazon EMR starts your cluster. 90 of your instances are available. It takes your cluster one hour to complete. How much will you be charged for this EMR usage for the first hour?**

A. $0.015

B. $1.50 per hour

C. $1.35 per hour ✓ **Correct**

D. $0

2. **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. ✓ **Correct**

Why is this correct?
Correct. Spot instances are cheaper than regular on-demand instances.

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**

3. **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	✓ **Correct**
D. S3 and Amazon Machine Learning	

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

5. 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 should you employ?

A. AWS Glue

B. EMR	✓ **Correct**
C. SQS	
D. Kinesis	✓ **Correct**

6. A large stuffed-animal maker is growing rapidly and constantly adding new animals to their product line. The COO wants to know customers reaction to each new animal, wants to ensure that their customers are enjoying the products, and use this information for future product ideas. The social media manager is tasked with capturing the feedback. After the data is ingested, the company wants to be able to analyze and classify the data in a cost-effective and practical way. A data scientist says she has already created an effective ML model in AWS for use if needed. How do you do this?

A. Use EMR and copy the raw data to Amazon S3. For long term analysis and historical reference, raw data is stored in Amazon S3.	
B. Use Redshift for processing and normalizing the data and requesting predictions from Amazon ML. Data is sent to Amazon SNS and delivered via email for further investigation.	
C. Use Lambda for processing and normalizing the data and requesting predictions from Amazon ML. Data is sent to Amazon SNS using Lambda and delivered via email for further investigation.	✓ **Correct**
D. Create Amazon Kinesis Streams and use one Kinesis stream to copy the raw data to Amazon S3. For long term analysis and historical reference, raw data is stored in Amazon S3.	✓ **Correct**

7. **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**

8. **You are designing the system backend for a big box store. Your data is going to have logs from every part of the business. The starting size of the data is 20 PB and spans about 8 years, and new data is going to be added as you go along. The logs do not conform to any specific schema. Which setup is most desirable for this scenario?**

A. Use Flume to migrate the data into HDFS and Hive/HiveQL on top of Hadoop to query the data. ✓ **Correct**
B. Use Kinesis Firehose to move the data into S3 and load into Redshift using `COPY` .

C. Move the data into S3 and load into Redshift using `COPY`.

D. Move the data into S3 and load into RDS using `COPY`.

9. **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 ✓ **Correct**

C. EMR ✓ **Correct**

D. Data Pipeline

10. **You need to create a recommendation engine for your e-commerce website that sells over 300 items. The items never change, and the new users need to be presented with the list of all 300 items in order of their interest. Which option do you use to accomplish this?**

A. Mahout ✓ **Correct**

Why is this correct?
Correct. This could be a good component of a custom solution.

B. Spark/SparkMLlib	✓ **Correct**
C. Amazon Machine Learning	✗ **Your Answer**

Why is this incorrect?
Incorrect. Amazon ML is limited to 100 "categorical' recommendations, so a custom system is required for this purpose.

D. RDS MySQL

11. 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	✓ **Correct**
D. Hadoop with Spark	

1. **Your client needs to load a 600 GB file into a Redshift cluster from S3, using the Redshift `COPY` command. The file has several known (and potentially some unknown) issues that will probably cause the load process to fail. How should the client most efficiently detect load errors without needing to perform cleanup if the load process fails?**

A. Split the 600 GB file into smaller 25 GB chunks and load each separately.
B. Compress the input file before running `COPY` .
C. Write a script to delete the data from the tables in case of errors.
D. Use the `COPY command` with the `NOLOAD` parameter. ✓ **Correct**

2. **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. Create a manual snapshot.
B. Use Lambda to create EBS snapshots.
C. Enable snapshot replication to another region and manually manage snapshots ✓ **Correct**
D. Enable cross-region replication in Redshift.

3. **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. ✓ **Correct**
D. Change the DISTSTYLE to EVEN.

4. **Your company deployed 100 sensors to measure traffic speeds on various highways that generated about 4 GB of data per month. The initial architecture used 400 GB RDS with EC2 instances. Over the next 3 months, there will be an additional 100,000 sensors added. You need to retain the data for at least 2 years for trends reporting. Which is the best solution to accomplish this?**

A. Replace the RDS instance with a 6 node Redshift cluster with 96 TB of storage. ✓ **Correct**
B. Write data from the sensors into a DynamoDB table and move the old data to a Redshift cluster.
C. Keep the current architecture, but upgrade RDS storage to 3 TB and 10 K provisioned IOPS.
D. Write data from the sensors into an SQS queue and then write into RDS.

5. **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	✓ **Correct**
B. Kafka	
C. Kinesis Firehose to S3 and Athena	
D. Spark Streaming with SparkSQL and MLlib	

6. **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	✓ **Correct**

7. You need to perform ad-hoc SQL queries on massive amounts of well-structured data. Additional data comes in constantly at a high velocity, and you don't want to have to manage the infrastructure processing it if possible. Which solution should you use?

A. Kinesis Firehose and RDS	
B. EMR running Apache Spark	
C. Kinesis Firehose and Redshift	✓ **Correct**
D. EMR using Hive	

8. Athena supports which file formats by default?

A. JSON	✓ **Correct**
B. PDF	
C. CSV	✓ **Correct**
D. Apache Parquet	✓ **Correct**

9. **You have large volumes of structured data in DynamoDB that you want to persist and query using standard SQL and your existing BI tools. What solution should you use?**

A. Use the `INSERT` command to load data in parallel directly to EMR from DynamoDB.
B. Use the `COPY` command to load data in parallel directly to Redshift. ✓ **Correct**
C. Use the `INSERT` command to load data in parallel directly to Redshift from DynamoDB.
D. Use the `COPY` command to load data in parallel directly to EMR from DynamoDB.

10. **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 platforms for your daily ad-hoc analysis. Additionally, you'll soon be developing new applications that require you to stream web application data in from multiple producers. Which tools enable you to do that?**

A. Write the data into SQS and dump the data into S3.
B. Create files on the producer and copy them to S3 and run EMR Hive to query the data.
C. Stream data from each producer into an S3 bucket and migrate the data nightly to RDS for reporting.
D. Use Amazon Kinesis to collect the data, use Kinesis Analytics for real-time analytics, and save the data in Redshift for trend analysis. ✓ **Correct**

11. **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` .	✓ **Correct**
B. Use `date_of_purchase` as the DISTKEY.	
C. Use LZO compression on the `date_of_purchase` column.	
D. Use `date_of_purchase` as the PARTITION KEY.	

AWS Big Data - Domain 5 - Visualization

1. **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. ✓ **Correct**

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.

2. 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 ✓ **Correct**

C. A stacked 100% area chart

D. A bubble chart

3. 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 ✓ **Correct**

B. QuickSight
C. Ganglia
D. YARN

4. **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	✓ **Correct**
D. Stacked Area Chart	

5. **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.	✓ **Correct**
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.	

6. You are using QuickSight to identify demand trends over multiple months for your top five product lines. Which type of visualization do you choose?

A. Scatter Plot	
B. Pie Chart	
C. Pivot Table	
D. Line Chart	✓ **Correct**

7. 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 ✓ **Correct**

C. QuickSight

D. Hue

8. **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 ✓ **Correct**

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.

B. A line chart

C. A stacked area chart

D. A histogram ✗ **Your Answer**

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.

9. 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**

10. You've been asked by the VP of People to showcase the current breakdown of the headcount for each department within your organization. What chart do you select to do this to make it easy to compare each department?

A. A line chart	
B. A column chart	
C. A pie chart	✓ **Correct**
D. A scatter plot	

11. **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	✓ **Correct**
D. Elasticsearch	

AWS Big Data - Domain 6 - Security

1. **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**

2. **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. ✓ **Correct**

3. **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. ✓ **Correct**

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.

B. Redshift should already have permissions after the cluster is created. Ask them to run a `COPY` command to load in the data. ✕ **Your Answer**

Why is this incorrect?
This isn't true. You will need to have some form of permissions to run the `COPY` command. Either through IAM Roles or API Keys.

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.

4. **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. ✓ **Correct**

C. Access the data through an Internet Gateway.

D. Access the data through a NAT Gateway.

5. **Your company stores very sensitive data on Redshift, which needs to be encrypted with keys that are fully controlled by your company. Which option should you use?**

A. AWS CloudHSM	✓ **Correct**
B. AWS KMS	
C. On-premise HSM	
D. S3-KMS	

6. **You are building an Amazon Redshift cluster within the shared services VPC of your organization. The cluster will host sensitive data. How can you control which networks can access the cluster?**

A. Run the cluster in a different VPC and connect through VPC peering.	
B. Only allow access to networks that connect with the shared services network via VPN.	
C. Create a database user inside the Amazon Redshift cluster only for users on the network.	
D. Define a cluster security group for the cluster that allows access from the allowed networks.	✓ **Correct**

7. **You have 30 customers, each of whom has a dedicated Kinesis Stream for 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. Submit a support request to Amazon inside the AWS console.	
B. Enable CloudWatch to monitor the streams.	
C. Move each customer into a separate AWS account and use consolidated billing.	
D. Tag the streams with the name of the customer.	✓ **Correct**

8. **What are the options to authenticate an IoT thing?**

A. KMS	
B. Amazon Cognito identities	✓ **Correct**
C. IAM groups and roles	✓ **Correct**
D. X.509 certificates	✓ **Correct**

9. **Server-side encryption is about data encryption at rest. That is, Amazon S3 encrypts your data at the object level as it writes it to disk in its data centers and decrypts it for you when you go to access it. There are a few options depending on how you choose to manage the encryption keys. One of the options is called 'Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3)'. Which of the following best describes how this encryption method works?**

A. A randomly generated data encryption key from Amazon S3, which is used by the client to encrypt the object data.
B. Each object is encrypted with a unique key employing strong encryption. As an additional safeguard, it encrypts the key itself with a master key that it regularly rotates. ✓ **Correct**
C. There are separate permissions for the use of an envelope key (that is, a key that protects your data's encryption key) that provides added protection against unauthorized access of your objects in S3 and also provides you with an audit trail of when your key was used and by whom.
D. You manage the encryption keys, and Amazon S3 manages the encryption, as it writes to disk, and decryption when you access your objects.

10. **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 ✓ **Correct**

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.