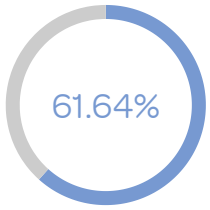

Great Start!

You did not pass the challenge on this attempt. This challenge is now locked and can be unlocked by using gems or by completing all of the recommended activities.

This challenge is now locked.

Unlock 0

Linux Academy

Go Back

Report Card

Expectations	Score
1. AWS Big Data - Domain 1 - Collection	72.73 %
2. AWS Big Data - Domain 2 - Storage	63.64 %
3. AWS Big Data - Domain 3 - Processing	36.36 %
4. AWS Big Data - Domain 4 - Analysis	63.64 %
5. AWS Big Data - Domain 5 - Visualization	63.64 %
6. AWS Big Data - Domain 6 - Security	80 %

Exam Breakdown

AWS Big Data - Domain 1 - Collection

1. **You have been hired as a consultant to provide a solution to integrate a client's on-premises data center to AWS. The customer requires a 300 Mbps dedicated, private connection to their VPC. Which AWS tool do you need?**

A. VPC peering ✕ **Your Answer** **Why is this incorrect?** Incorrect. VPC peering will only help if the client has multiple VPCs that they need to connect.
B. Data Pipeline
C. Direct Connect ✓ **Correct** **Why is this correct?** Correct! Direct Connect will provide a dedicated and private connection to an AWS VPC.
D. EMR

2. **Your company has two batch processing applications that consume financial data about the day's stock transactions. Each transaction needs to be stored durably and guarantee that a record of each application is delivered so the audit and billing batch processing applications can process the data. However, the two applications run separately and several hours apart and need**

access to the same transaction information. After reviewing the transaction information for the day, the information no longer needs to be stored. What is the best way to architect this application?

A. Use SQS for storing the transaction messages; when the billing batch process performs first and consumes the message, write the code in a way that does not remove the message after consumed, so it is available for the audit application several hours later. The audit application can consume the SQS message and remove it from the queue when completed.

✕ **Your Answer**

Why is this incorrect?
Incorrect. SQS would make this more difficult because the data does not need to persist after a full day.

B. Use Kinesis to store the transaction information. The billing application will consume data from the stream and the audit application can consume the same data several hours later.

✓ **Correct**

Why is this correct?
Correct. Kinesis appears to be the best solution that allows multiple consumers to easily interact with the records.

C. Store the transaction information in a DynamoDB table. The billing application can read the rows while the audit application will read the rows then remove the data.

D. Use SQS for storing the transaction messages. When the billing batch process consumes each message, have the application create an identical message and place it in a different SQS for the audit application to use several hours later.

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. **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.

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 RDS Aurora	
B. AWS Data Pipeline	
C. AWS Data Migration Service	✓ **Correct**
D. AWS Kinesis	

6. **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.	

7.

You need to analyze clickstream data on your website from multiple applications. You want to analyze the pattern of pages a consumer clicks on and in what order. You need to be able to use the data in real time and want to manage as little infrastructure as possible. Which option would meet this requirement?

A. Publish web clicks by session to an Amazon SQS.
B. Use ElasticMap Reduce to ingest the data and analyze it.
C. Send click events directly to Amazon Redshift and then analyze them with SQL.
D. Use Amazon Kinesis with a worker to process the data received from the Kinesis stream. ✓ **Correct**

8. Your organization has a variety of different services deployed on EC2 and needs to efficiently send application logs over to a central system for processing and analysis. They've determined it is best to use a managed AWS service to transfer their data from the EC2 instances into Amazon S3 and they've decided to use a solution that will do what?

A. Installs the AWS Direct Connect client on all EC2 instances and uses it to stream the data directly to S3.
B. Leverages the Kinesis Agent to send data to Kinesis Data Streams and output that data in S3.
C. Ingests the data directly from S3 by configuring regular Amazon Snowball transactions.
D. Leverages the Kinesis Agent to send data to Kinesis Firehose and output that data in S3. ✓ **Correct**

9. **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.

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** **Why is this correct?** Correct! This architecture satisfies the requirements of the scenario.

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. ✕ **Your Answer**

Why is this incorrect?
Incorrect. While Direct Connect is certainly part of the solution, there are still missing components of this architecture; a VPN and Direct Connect partner request are missing from the requirements.

11. You have 3 Kinesis KCL applications that are reading a Kinesis Streams stream and are falling behind. CloudWatch is emitting `ProvisionedThroughputExceededException` errors on your stream. What corrective active do you need to take to make sure you can use at least 3 KCL applications?

A. Convert the Kinesis Streams to Kinesis Firehouse.

B. Pick a different partition key for your stream.

C. Add more shards to your Kinesis Stream. ✓ **Correct**

D. Add more KCL applications.

AWS Big Data - Domain 2 - Storage

1. **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.	✓ **Correct**
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.	

2. **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.	✓ **Correct**
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.	

3. **An application saves log data to S3. Your development team wants to keep logs for one month for troubleshooting purposes, and then purge them. What can you do to support this?**

A. Configure the lifecycle configuration rules on the S3 bucket.	✓ **Correct**
B. Add a bucket policy on the S3 bucket.	
C. Create an IAM policy for the S3 bucket.	
D. Enable CORS on the S3 bucket.	

4. **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.

5. You work for a social media start-up and need to analyze the effectiveness of your new marketing campaign from your previous one. Which process should you use to record the social media replies in a durable data store that can be accessed at any time for analytics of historical data?

A. Read the data from the social media sites, store it in DynamoDB, and use Apache Hive with Amazon Elastic MapReduce for analytics. ✓ **Correct**

Why is this correct?
Correct. DynamoDB is an effective data store for this data that allows key-value-based querying, but for additional analytics, you can add a layer on top of that with Apache Hive and EMR.

https://aws.amazon.com/blogs/aws/aws-howto-using-amazon-elastic-mapreduce-with-dynamodb/ (https://aws.amazon.com/blogs/aws/aws-howto-using-amazon-elastic-mapreduce-with-dynamodb/)

B. Read the data from the social media sites, store it in Amazon Glacier, and use AWS Data Pipeline to publish it to Amazon Redshift for analytics. ✗ **Your Answer**

Why is this incorrect?
Incorrect. Glacier is not a good solution for anything other than archival storage.

C. Read the data from the social media sites, store it with Amazon Elastic Block Store, and use AWS Data Pipeline with Amazon Kinesis for analytics.

D. Read the data from the social media site, store it with Amazon Elastic Block store, and use Amazon Kinesis Data Streams for analytics.

6. **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. ✕ **Your Answer**

Why is this incorrect?
Incorrect. You can't change the primary key on a table. You would have to recreate the table to do this.

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**

Why is this correct?
Correct. Global secondary indexes are particularly useful for tracking relationships between attributes that have a lot of different values. For example, you could create a DynamoDB table with `CustomerID` as the primary partition key for the table and `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. Using the primary key, you could quickly get the record for any customer. Using the global secondary index, you could efficiently query for all customers that live in a given ZIP code.

Further Reading https://aws.amazon.com/dynamodb/faqs/ (https://aws.amazon.com/dynamodb/faqs/)

7.

You're working with a new client that is currently building out a new image-based website. This client currently has 23,000 users worldwide uploading an average of five images per day, each about 1MB each. The application stores these images on-premises but the client is concerned that with an increase in user growth and interactions, they will need a substantial amount of additional space beyond their current capacity and is hoping to migrate their application to the cloud. The application requires that images are always available for users to view or download as quickly as possible, but in the future, the developers are considering adding features to allow users to pay less but wait longer to retrieve family image albums or simply put a cap on the number of images they load. What components of the architecture do you suggest they use when creating the image-hosting portion of the application?

A. S3 should be used to store images and serves as a CloudFront origin. CloudFront should be used to cache images to users and serve them up to users more rapidly.	✓ **Correct**
B. DynamoDB should be used to store images in combination with the DynamoDB Accelerator (DAX) to serve the images up with sub-millisecond latency to users at edge locations.	
C. Store the images within EFS and maintain a fleet of t2.micro instances in multiple AWS regions to serve the images quickly to users across the world.	
D. Migrate existing application code to ECS to allow them to scale out the compute capacity required, then migrate the images to EFS using the Storage Gateway.	

8. **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 vault lock policy for archives less than seven years old. ✓ **Correct**
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.

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

A. Local Secondary Index
B. Global Secondary Index ✓ **Correct**
C. None of these.
D. Primary Hash Index

10.

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. ✕ **Your Answer**

Why is this incorrect?
Incorrect. This probably isn't the issue, 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.

D. Review the DynamoDB partition keys and determine how you can efficiently randomize them. Then, if necessary, increase read capacity units. ✓ **Correct**

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.

11.

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.	✓ **Correct**
B. Create an LSI on location.	✗ **Your Answer**
Why is this incorrect? Incorrect. LSIs cannot be created after table creation.	
C. Increase the RCUs.	✓ **Correct**
D. Increase the WCUs.	✓ **Correct**
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.	

AWS Big Data - Domain 3 - Processing

1. **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**

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.

D. Data Pipeline

2. **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 ✗ **Your Answer**

Why is this incorrect?
Incorrect. Close, but you'll save a little bit of money because you aren't charged for those 10 instances that haven't started yet.

C. $1.35 per hour ✓ **Correct**

Why is this correct?

Correct. Billing commences when Amazon EMR starts running your cluster. You are only charged for the resources consumed. For example, let's say you launched 100 Amazon EC2 Standard Small instances for an Amazon EMR cluster, where the Amazon EMR cost is an incremental $0.015 per hour. The Amazon EC2 instances will begin booting immediately, but they won't necessarily all start at the same moment. Amazon EMR will track when each instance starts and will check it into the cluster so that it can accept processing tasks. In the first 10 minutes after your launch request, Amazon EMR either starts your cluster (if all your instances are available) or checks in as many instances as possible. Once the 10-minute mark has passed, Amazon EMR will start processing (and charging for) your cluster as soon as 90% of your requested instances are available. As the remaining 10% of your requested instances check in, Amazon EMR starts charging for those instances as well. So, in the above example, if all 100 of your requested instances are available 10 minutes after you kick off a launch request, you'll be charged $1.50 per hour (100 *$0.015) for as long as the cluster takes to complete. If only 90 of your requested instances were available at the 10-minute mark, you'd be charged $1.35 per hour (90* $0.015) for as long as this was the number of instances running your cluster. When the remaining 10 instances check in, you'd be charged $1.50 per hour (100 * $0.015) for as long as the balance of the cluster takes to complete.

Further Reading https://aws.amazon.com/emr/faqs/ (https://aws.amazon.com/emr/faqs/)

D. $0

3. 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**

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.

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. ✓ **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.

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. ✓ **Correct**

6. **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 ✓ **Correct**

D. Regression

7. **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**

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)

C. SQS

D. Kinesis ✓ **Correct**

8.

You get daily dumps of transaction data into S3 which is batch processed into EMR on a nightly basis. The size of the data spikes up and down regularly. What can be done to reduce the processing time?

A. Add more orchestration nodes to the cluster.
B. Use Spot instances for the nodes.
C. Add task nodes using "based on CPU" metrics from Ganglia. ✓ **Correct**
D. Process the data in Redshift instead.

9. 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 ✓ **Correct** **Why is this correct?** Correct. Spark has micro-batching but can guarantee only-once-delivery if configured correctly.
C. SQS standard queues ✗ **Your Answer** **Why is this incorrect?** Incorrect. SQS standard queues can't guarantee only-once delivery.
D. None of these

10. You have a Kinesis stream with four shards that receive data from various IoT devices. There is a lambda transformation function attached to the streams that fan out the data to eight destinations. How many total lambda functions get invoked concurrently with the shards?

A. 8

B. 32

C. 4 ✓ **Correct**

Why is this correct?
Correct. You will need one concurrent function per shard.

D. 1 ✗ **Your Answer**

Why is this incorrect?
Incorrect. There is more than one shard and you will need one concurrent function per shard.

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**

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

D. Hadoop with Spark ✕ **Your Answer**

Why is this incorrect?
Incorrect. Spark and Hadoop are great components of many big data architectures, but they aren't specifically used for key-value storage.

AWS Big Data - Domain 4 - Analysis

1. 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.

2. **Your data warehouse is running on Redshift. On average, you have 5 users that are logged in at any given time and running queries. You have a query that runs for 50 minutes. While you run your query, any new queries that are initiated by anyone else must wait for at least 50 minutes before returning data; normally, their queries come back within a minute. Which of the following can get the queries running faster for everyone else?**

A. Put your query into a separate WLM queue with concurrency set to 1.	✓ **Correct**
B. Add more nodes to the Redshift cluster.	
C. Create a separate WLM queue for everyone else's queries with concurrency set to 3.	✓ **Correct**
D. Temporarily up the concurrency by one.	✓ **Correct**

3. **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**

Why is this correct?
Creating EBS snapshots for Redshift is not possible. Manual snapshots would be in the same region and automatic cross-region replication for Redshift is not a current feature.

D. Enable cross-region replication in Redshift. ✕ **Your Answer**

Why is this incorrect?
Automatic cross-region replication for Redshift is not a current feature.

4. You need to perform ad-hoc SQL queries on structured data that will be generated by a fleet of IoT devices about every 20 minutes. What services should you use if you want to be able to also compare this incoming data with a massive amount of existing data inside one centralized data warehouse?

A. EMR + Redshift ✕ **Your Answer**

Why is this incorrect?
EMR can process data if we need it to, but we already have the data structured the way we want.

B. Kinesis Firehose and Redshift ✓ **Correct**

Why is this correct?
Kinesis Firehose provides a managed service for aggregating streaming data and inserting it into Redshift. Redshift also supports ad-hoc queries over well-structured data using an SQL-compliant wire protocol, so the business team should be able to adopt this system easily.

Further Reading http://docs.aws.amazon.com/redshift/latest/dg/c_redshift-sql.html (http://docs.aws.amazon.com/redshift/latest/dg/c_redshift-sql.html)

C. EMR using Hive

D. Kinesis Data Streams sending data into RDS

5. **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 ✕ **Your Answer**

Why is this incorrect?
EMR and Spark (using Spark Streaming for example) could work in this scenario if there weren't an explicit request not to have to manage the infrastructure that does the processing.

C. Kinesis Firehose and Redshift ✓ **Correct**

Why is this correct?
Amazon Kinesis Firehose is the easiest way to ingest streaming data into Amazon Redshift. Amazon Kinesis Firehose automatically batches and compresses data before loading it into Amazon Redshift and makes the data available for processing and analytics.

D. EMR using Hive

6. **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 ✓ **Correct**
D. RDS, EMR, Data Pipeline, QuickSight

7. **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**

8. **Which tool allows you to search through CloudWatch logs?**

A. Elastic Map Reduce	
B. Elasticsearch	✓ **Correct**
C. ElastiCache	
D. Elastic Cloud Computing	

9. 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**

10.

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`.	✕ **Your Answer**
Why is this incorrect? S3 `COPY` is the fastest loading mechanism of data from S3 to Redshift, so this isn't a good option.	
B. Store the data already sorted in the sortkey order.	✓ **Correct**
Why is this correct? Loading data already in the sortkey order will allow Redshift to save time and not need to sort data.	
C. Copy the data to S3 and use `COPY` to move the data into Redshift.	✓ **Correct**
D. Compress the data inside of S3 before loading it into Redshift.	✓ **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 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**

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

C. A stacked 100% area chart

D. A bubble chart — ✗ **Your Answer**

Why is this incorrect?
Bubble charts are useful when comparing many different datasets and trying to ascertain some correlation.

2. **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.	✓ **Correct**
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.	

3. **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**
B. A line chart	
C. A stacked area chart	
D. A histogram	

4. **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.	

5. 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**

6.

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	

7. 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	✗ Your Answer
Why is this incorrect? Column charts are great for showcasing trends over time. They might work here, but since we want to know the percentages of a whole, they aren't the best solution.	
C. A pie chart	✓ Correct
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.

D. A scatter plot

8. **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**

9. **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**

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!

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**

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

10. 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**

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.

C. QuickSight ✕ **Your Answer**

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.

D. Hue

11. 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

AWS Big Data - Domain 6 - Security

1. 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**

2. 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. ✓ **Correct**

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.

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. ✗ **Your Answer**

Why is this incorrect?
While this would work, it doesn't meet the requirements as key management can perform in the cloud instead of on-premises.

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.

3. **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**

4. **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.

5. 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**

6. **You've just deployed a Redshift cluster and are attempting to connect to it using Postico. You've already managed to connect to a separate Redshift cluster for another client that you're working with, but now you're unable to connect to this new cluster. What might be the issue?**

A. You need to install the ODBC or JDBC drivers to connect.

B. You need to provision an SSH to allow yourself to sign into the instance from your terminal instead.

C. You need to reconfigure your VPC to allow SSH traffic in on port 22 from your IP.

D. You need to reconfigure your VPC to allow TCP traffic in on port 5439 from your IP. ✓ **Correct**

7. **You need to alert your administrators every time downloads of specific objects in an S3 bucket occur. How do you do this?**

A. Create a CloudWatch metric for BucketGetRequests and alert off that using SQS.

B. Create a DynamoDB stream that logs the access requests and have a Lambda function monitoring that for GET requests in order to send SMS via SNS when appropriate. ✗ **Your Answer**

Why is this incorrect?

DynamoDB streams are not configurable to log access requests to S3 buckets. While those GET requests could potentially trigger Lambda, it isn't mentioned how that would happen in this answer.

C. Create a Lambda function to trigger of S3 Bucket PUT object requests and send an email via SES to your administrators.

D. Create a CloudTrail Trail that integrates with CloudWatch metrics and SNS to send alerts via SMS. ✓ **Correct**

Why is this correct?
CloudTrail is the best tool to monitor API-level actions inside of an AWS account. Because of how it integrates with CloudWatch metrics and SNS for text messages, it makes the best solution in this case.

8. **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**

9.

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	

10. 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**