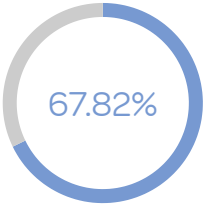

Great Start!

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	63.64 %
2. AWS Big Data - Domain 2 - Storage	63.64 %
3. AWS Big Data - Domain 3 - Processing	72.73 %
4. AWS Big Data - Domain 4 - Analysis	72.73 %
5. AWS Big Data - Domain 5 - Visualization	72.73 %
6. AWS Big Data - Domain 6 - Security	60 %

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

A. Log all events using the Kinesis Producer Library.

B. Log critical events using the Kinesis Producer Library, and log informational events using the PutRecords API method. ✕ **Your Answer**

Why is this incorrect?
Using the Kinesis Producer library wouldn't delay processing in an application which is required for the critical events.

C. Log critical events using the PutRecords API method, and log informational events using the Kinesis Producer Library. ✓ **Correct**

Why is this correct?
The PutRecords API can be used in code to be synchronous; it will wait for the API request to complete before the application continues. This means you can use it when you need to wait for the critical events to finish logging before continuing. The Kinesis Producer Library is asynchronous and can send many messages withing needing to slow down your application. This makes the KPL ideal for the sending of many non-critical alerts asynchronously.

D. Log all events using the PutRecords API method.

3. **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 ✓ **Correct**

VPN connection to one or more Direct Connect connections as needed.

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.

4. **Your IoT application has smoke sensors in various hotels. You need to collect this data in real-time and log it all to S3 and, in the event a sensor detects smoke, to send out an alert. What steps do you need to take to accomplish this?**

A. Create a rule to send a push notification to all users using Amazon SNS. ✕ **Your Answer**

Why is this incorrect?
A rule would be needed, but not for this purpose you would instead want it to filter out the smoke sensors that failed.

B. Create a rule to filter the smoke sensors that detect smoke. ✓ **Correct**

Why is this correct?
Correct. You need to create a rule to filter the smoke sensors and an action to push the notification out to SNS.

C. Create an action to filter the smoke sensors that detect smoke.

D. Create an action to send a push notification to all users using Amazon SNS. ✓ **Correct**

5. Which of the following statements are true about Apache Kafka? (Choose two)

A. Apache Kafka stores streaming data records in an ordered, fault-tolerant way. ✓ **Correct**

B. Apache Kafka acts as a drop in replacement for the NGINX web server.

C. Apache Kafka stores streaming data records in fault-tolerant way, but does not guarantee the order of the records.

D. Apache Kafka provides a buffer between producers of the data and consumers of the data. ✓ **Correct**

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

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

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

C. Submit system and application logs directly to Amazon Kinesis Streams using the Kinesis agent on the ✓ **Correct**

front-end and application machines themselves.

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

7. **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.

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**

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.

8. **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.

9. **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
B. Data Pipeline
C. Direct Connect ✓ **Correct**
D. EMR

10. **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 ✕ **Your Answer**

Why is this incorrect?
Incorrect. Spark Streaming does it in batches.

D. Kinesis Stream ✓ **Correct**

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

11. **Your application generates a 1 KB JSON payload that needs to be queued and delivered to EC2 instances for applications. At the end of the day, the application needs to replay the data for the past 24 hours. In the near future, you also need the ability for other multiple EC2 applications to consume the same stream concurrently. What is the best solution for this?**

A. Kinesis Data Streams ✓ **Correct**

B. Kinesis Firehose
C. SNS
D. SQS

AWS Big Data - Domain 2 - Storage

1. **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.	

2.

You have an application that is currently in the development stage but is expected to write 2,400 items per minute to a DynamoDB table; each 2Kb in size or less and then fluctuate to 4,800 writes of items (of the same size) per minute on weekends. There may be other fluctuations within that range in the future as the application develops. How should this table be created? It is important to the success of the application that the vast majority of user requests are met.

A. Provision a base WCU of 80 and then schedule regular increases to 160 WCUs when a higher load is expected.
B. Set up an auto-scaling policy on the DynamoDB table that doesn't let the traffic dip below the usual load and allows it to scale to meet demand. ✓ **Correct**
C. Enabled DynamoDB streams have a Lambda function triggered to review the current capacity on each change to the table.
D. Provision a base WCU of 160 and then schedule a job that adds 160 more WCUs when a higher load is expected.

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. **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	

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

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

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

B. Convert the file format to CSV format.

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

Why is this incorrect?
While splitting the file into smaller files is an effective way to improve performance, you could actually split the file down even more and gain more performance given that each node of a cluster has multiple slices that can each load data in parallel.

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

6. You need to leverage Amazon Simple Storage Service (S3) for backups using third-party software. How can you ensure the credentials provisioned for the third-party software limit access to the `backups-18` folder of the `abc-corp-backups-2018` bucket to allow that third party to send data into the bucket from another AWS account?

A. A custom folder ACL on the `backups-18` folder. ✕ **Your Answer**

Why is this incorrect?
Incorrect. ACLs can be applied at the bucket or object level, not at the folder level.

B. Provide the third-party software an access key tied to an IAM user of a developer with access to that bucket already.

C. A custom bucket policy on the `abc-corp-backups-2018` bucket limited to the Amazon S3 API in the `backups-18` folder. Specify the database user inside of the other AWS account that you want to have access.

D. A custom bucket policy on the `abc-corp-backups-2018` bucket limited to the Amazon S3 API in the `backups-18` folder. Specify the principal of the other AWS account. ✓ **Correct**

Why is this correct?
Correct. This could work to allow the external account access.

7. **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. **Why is this incorrect?** Incorrect. Using a partition key with multiple data fields connected in this way will massively reduce performance.	✕ **Your Answer**
B. Use one table for each week with a partition key that is the sensor ID and a key for the timestamp. **Why is this correct?** Correct. This would allow the look up sensor data within a particular time range.	✓ **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.	

8. **You have a web application that allows customers to upload orders to an S3 bucket. The S3 bucket has an S3 upload event trigger a Lambda function that inserts a message to an SQS queue. Then, an auto-scaling group of t2.medium EC2 instances reads messages from the**

queue, processes them, and stores them in a DynamoDB table partitioned by unique order ID and provisioned to meet current capacity with 15 WCUs and 15 RCUs.

Traffic for this application is expected to increase by 500% next month. What component of this architecture might need to be revisited?

A. The SQS queue will need to be turned into a FIFO queue to support the additional load.
B. The DynamoDB table will need additional provisioned capacity or to be set up with auto-scaling. ✓ **Correct**
C. The Lambda function will require additional capacity requested via AWS support.
D. The EC2 instances will need to be upgraded to a larger type to support the load.

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

A. 40 ✓ **Correct** **Why is this correct?** Correct. When doing the calculation, here's what you would do: 1.5 KB / 1 = 1.5 KB, which gets rounded up to 2. 2 RCUs * 20 items = 40 WCUs Remember that Writes Capacity Units support one 1 KB-sized item write per second. You also need to round up item sizes to the nearest 1 KB.
B. 80
C. 10

D. 20 ✕ **Your Answer**

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

10. 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

11. 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`.	

AWS Big Data - Domain 3 - Processing

1. You need to be able to access resources in S3 and then write data to tables in S3. You also need to be able to load table partitions automatically from Amazon S3. Which Big Data tool enables you to do so?

A. EMR, Hive, and Redshift Spectrum. ✓ **Correct**

Why is this correct?

Correct. Hive allows user extensions via user-defined functions written in Java. Amazon EMR has made numerous improvements to Hive, including direct integration with DynamoDB and Amazon S3. For example, with Amazon EMR you can load table partitions automatically from Amazon S3, you can write data to tables in Amazon S3 without using temporary files, and you can access resources in Amazon S3, such as scripts for custom maps and/or reduce operations and additional libraries. See https://docs.aws.amazon.com/redshift/latest/dg/c-using-spectrum.html (https://docs.aws.amazon.com/redshift/latest/dg/c-using-spectrum.html) for Redshift Spectrum information.

https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf (https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf)

B. Redshift and Athena ✕ **Your Answer**

Why is this incorrect?
Incorrect. Redshift and Athena are great and will let you query S3 data using SQL, either directly in S3 by using Athena or after loading into Redshift with a `COPY` command, but neither enables you to load table partitions automatically from S3.

C. Redshift and SQL

D. EMR and Pig

2. 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.

3.

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. ✕ **Your Answer**

Why is this incorrect?
Incorrect. EMR isn't a great solution to collect the data initially.

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**

Why is this correct?
Correct. This is a great part of the solution. It can collect and store the data long-term for future use.

4. **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 ✗ **Your Answer**

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

5. **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

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

A. Amazon Data Pipeline	
B. Amazon EMR with Spark	✓ **Correct**
C. Amazon Redshift	

D. Amazon Machine Learning

8. **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	

9. **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**
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**

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**
D. 1	

11. What are some of the benefits of running Spark vs. MapReduce?

A. Regardless of data size, Spark is always more cost-efficient.	
B. Spark eliminates the complexities and requirements of running map and reduce operations.	
C. Spark can use in-memory processing to speed up queries.	✓ **Correct**
D. Spark has an optimized directed acyclic graph (DAG) execution engine.	✓ **Correct**

AWS Big Data - Domain 4 - Analysis

1. **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.	

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 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**

4. Athena supports which file formats by default?

A. JSON	✓ **Correct**
B. PDF	

C. CSV	✓ **Correct**
D. Apache Parquet	✓ **Correct**

5. **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**

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	✗ **Your Answer**

Why is this incorrect?

EMR might work in this case, but the type of data appears to lend itself better to tools that are specifically designed to query very heterogeneous data like this.

B. Redshift

C. QuickSight

D. Elasticsearch ✓ **Correct**

Why is this correct?

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

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

7. **A client comes to you and requests access to a specific Redshift compute node that they want access to in order to add performance monitoring software. They plan to install open source software on this node in order to have access to data such as disk utilization and query performance. How do you address the request?**

A. You change the security settings in the VPC for Redshift to allow ingress traffic and provide them with an SSH key. ✗ **Your Answer**

Why is this incorrect?
You actually can't connect to Redshift with SSH!

B. You inform them that directly connecting is not possible, but that the AWS console and CloudWatch provide many of the metrics. ✓ **Correct**

Why is this correct?
Because Redshift is a somewhat managed service, you can't access the underlying operating system. But it also means that AWS is nice and provides you metrics in CloudWatch to help manage your cluster.

C. You use the Amazon Redshift APIs to install the performance monitoring software for them.

D. You provide them with a unique SSH private key and add the public key to the cluster's acceptable keys list so they can access the underlying EC2 Instance.

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

A. SQS

B. EMR

C. Athena ✓ **Correct**

D. Redshift

9. **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	

10. **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.	✗ **Your Answer**
Why is this incorrect? While this is another valid distribution style, it's not the most effective one in the situation described.	
C. Change the DISTSTYLE to ALL.	✓ **Correct**
Why is this correct?	

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

D. Change the DISTSTYLE to EVEN.

11. 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

AWS Big Data - Domain 5 - Visualization

1. 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	

2. **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.	

3. **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 ✕ **Your Answer**

Why is this incorrect?
Stacked area charts are useful when you want to evaluate how different segments of data changed in respect to one another over time. That's not entirely what we want here.

D. A bubble chart

4. **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**

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

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

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

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

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

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 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.	

8. **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	

9. 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	

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**

C. QuickSight

D. Hue

11. 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 ✗ **Your Answer**

Why is this incorrect?
Line charts are useful when we have many different groupings of data, and we want to compare them over time. In this case, we're comparing data within the same period, so this isn't the best choice.

C. A stacked area chart

D. A histogram

AWS Big Data - Domain 6 - Security

1. **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**
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.	

2. **You have to design an EMR system where you will be processing highly confidential data. What can you do to ensure encryption of data at rest?**

A. SSE-KMS	✓ **Correct**

B. VPN

C. TLS

D. LUKS ✓ **Correct**

Why is this correct?
Linux Unified Key Setup, or LUKS, is a disk encryption specification that can help encrypt sensitive data at rest.

3. What are the options to authenticate an IoT thing?

A. KMS

B. Amazon Cognito identities ✓ **Correct**

C. IAM groups and roles ✓ **Correct**

Why is this correct?
IAM groups, roles, and policies can be used in combination to help authenticate IoT devices by associating them with the appropriate permissions.

D. X.509 certificates ✓ **Correct**

Why is this correct?
You can store X.509 certificates on a device and use it when sending date to authenticate that the device is not forging data. This is similar to how SSH keys work.

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

A. The mobile app should call a REST-based service that stores data on Amazon EBS. Deploy the service on multiple EC2 instances across two Availability Zones.
B. The mobile app should authenticate with an Amazon Cognito identity that is authorized to write to an Amazon Kinesis Firehose with an Amazon S3 destination. ✓ **Correct**

C. The mobile app should write to an S3 bucket that allows anonymous PutObject calls.

D. The mobile app should authenticate with an embedded IAM access key that is authorized to write to an Amazon Kinesis Firehose with an Amazon S3 destination.

6. **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. ✕ **Your Answer**

Why is this incorrect?
You cannot access S3 through a VPN in a way that avoids these concerns.

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

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

C. Access the data through an Internet Gateway.

D. Access the data through a NAT Gateway.

7. **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**

8. **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**

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**

Why is this correct?
CloudHSM is a physical device that is attached to your VPC by AWS giving you full and exclusive control to the HSM, and only you have access/control of the keys.

B. AWS KMS

C. On-premise HSM — ✗ **Your Answer**

Why is this incorrect?
While you could have your Hardware security module or HSM, the CloudHSM option provided by AWS still gives you full and exclusive control to the HSM - only you have access/control of the keys.

D. S3-KMS

10. **What is the result of the following bucket policy?**

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

A. It will deny all actions if the object prefix is `data_team_` .	
B. It will allow all actions if the object is in the finance subdirectory of `mybucket` .`\`	
C. It will allow all actions only against objects with the prefix `data_team_` in the `mybucket` bucket.	✓ **Correct**
D. It allows access to objects in the `data_team_` bucket namespace.	