1. Sending a lot of 100 B data records and want to ensure that Kinesis receives data:
    a. Use KPL for its asynchronous features and to ensure optimal throughput
    b. Achieve maximum throughput through batching collection and aggregation
2. Collecting log files in mass from Kinux servers running on premise and need a retry mechanism embedded and monitoring through CW. The logs should end up in Kinesis: Use Kinesis Agent
3. Consuming from a Kinesis Stream with 10 shards that recipes on average 8 MB/second of data from various producers using KPL. It is observed through CW metrics that the throughput is 2 MB/second so the app is lagging:
    a. Increase RCU/WCU because the most likely cause is that DynamoDB is underprovisioned so checkpointing does not append fast enough and results in a lower throughput for KCL based applications
4. Need a managed service that deelivers data to S3, scales automatically, billed only for actual usage of the service and be able to handle peak loads: Use Kinesis Firehose
5. Collecting data from IoT devices at scale and want to forward it into Firehose:
    a. Send that data into an IoT topics and define a rule action
6. A DX connection is set up on one location to ensure traffic into AWS is going over a private network but want to set up a failover connection that is reliable and redundant (cannot afford it being down for too long):
    a. Use another site to site VPN as a backup connection
    b. Not as private as another DX setup but it is definitely more reliable since it leverages public web
7. To store 8 TBs of usable storage, use Snowcone when Snowball doesn't fit (like space constrained places)
8. An app on EC2 creates images after photos are uploaded to S3: These images only need to be kept for 45 days. The source images should be immediately retrieved for 4 days and afterwards the user can wait up to 6 hours:
    a. S3 images can be stored on Standard, with lifecycle configuration to transition to Glacier after 45 days
9. An app has a lot of files from on premise NFS storage being inserted into S3. For the data integrity verification, the app downloads the files right after upload:
    a. App receives a 200 because S3 for new PUT is strongly consistent
10. Want to be able to recover deleted S3 objects immediately for 1 day, although this may happen rarely. After this time and for up to 364 days, deleted objects should be recoverable within 48 hours:
    a. Enable S3 Versioning for object versions: deleted objects are hidden by a delete marker but can be recovered
    b. Transition these no current version of the objects to S3 IA and after to Deep Archive
11. Gathering various files to analyze them once a month using Athena. Must return query results immediately, want to minimize risk of losing files and want to minimize costs:
    a. Use S3 Infrequent Access
12. Must archive all logs created by apps and ensure they cannot be modified or deleted for at least 7 years:
    a. Use Glacier with a vault lock policy
13. Generating thumbnails in S3 from images but some in the prefix are rarely read so can optimize costs by moving them to another S3 tier. For the least amount of changes:
    a. Create a lifecycle rule for this prefix
14. An app plans to have 15,000 R/Ws per second to S3 from thousands of device IDs:
    a. Use convention device-id/yyyy-mm-dd
15. Files should be encrypted in S3 but don't want to manage the encryption yourself. Want to have control over the encryption keys and ensure they are securely stored in AWS:
    a. Use SSE-KMS
16. Website is deployed and sources its images from an S3 bucket: all works fine on the Internet but when start website locally for dev, the images are not getting loaded:
    a. The issue is S3 CORS
17. An app is taking files from local on-premise NFS storage and inserting them to S3. As part of the data integrity verification, want to ensure files have been properly uploaded at minimal cost:
    a. Proceed by computing the local ETag for ach file and comparing them with AWS S3 ETag
18. An app wants to give users access to their own personal space in S3:
    a. Use Cognito Identity Federation
19. Company needs to aggregate daily stock data from exchanges into a data store. Requires that data is streamed directly into the data store but occasionally allows data to be modified using SQL. Solution should integrate complex, analytic queries running with minimal latency and provide a BI dashboard for visualization:
    a. Use Kinesis Firehose to stream data into Redshift and use Redshift as a data source of QS for the dashboard
20. Company hosts a data lake in S3 and a data warehouse in Redshift and uses QS to build dashboards. Want secure access from on-premise Active Directory to QuickSight:
    a. Use VPC Endpoint to connect to S3 from QS and IAM role to authenticate Redshift
21. Company hosts an app on AWS and new features are released weekly. A solution must be deployed that analyzes logs from each EC2 instance to ensure the app is working as expected after each deployment. The solution should be highly available with the ability to display new information the minimal delays:
    a. Use CW subscriptions to get access to a real time feed of logs and have the logs delivered to Kinesis Data Streams to further ush the data to Opensearch and Opensearch dashboards

22. A critical app uses Apache HBase in EMR, which is configured on a single master node. The company has over 5TB of data stored in HDFS and wants a cost effective solution to make HBase data highly available.
    a. Store data on EMRFS instead of HDFS and enable EMRFS consistent view
    b. Create a primacy EMR HBase cluster with multiple master nodes
    c. Create a secondary EMR HBase read replica cluster in a separate AZ
    d. Point both clusters to the same HBase root directly in the same S3 bucket
23. Glue is used to organize, clean, validate, format 200 GB dataset. A job is triggered to run with Standard worker type. After 3 hours, the Glue job status is still running but logs from the job run show no error codes. To improve the job execution time without overprovisioning:
    a. Enable job metrics in Glue to estimate the number of DPUs and based on the profiled metrics, increase the value of the maxim capacity job parameter
24. Company uploads CSV files to S3 and a Glue Crawler does the discovery to create tables/schemas. A Glue job writes processed data from created tables to Redshift database and handles column mapping and creating the Redshift table. When the job is rerun for any reason in a day, duplicate records are introduced to the Redshift table. Want a solution that updates the Redshift table without duplicates when jobs are rerun:
    a. Modify the Glue job to copy the rows into a staging table and add SQL commands to replace the existing rows in the main table as post actions in the DynamicFrame Writer class
25. A streaming app is reading data from Kinesis Data Streams and immediately writing data to an S3 bucket every 10 seconds. The app is reading data from hundreds of shards and batch intervals cannot be changed due to a separate requirement. The data is accessed by Athena but users are seeing a degradation in query performance as time progresses. To improve query performance:
    a. Merge the files in S3 to form larger files
26. Company uses OpenSearch to store, analyze clickstream data. The company ingests 1 TB of data daily using Kinesis Firehose and stores 1 days worth of data in an Elasticsearch cluster. The company has slow query performance on ElasticSearch index and occasionally sees errors from Firehose when attempting to write to the index. The Elasticsearch cluster has 10 nodes running in a single index and 3 dedicated master nodes. Each data node has 1.5 TB of EBS storage attached and the cluster is configured with 1000 shards. Occasionally, JVM Memory Pressure errors are found in the cluster logs. To improve the performance of ElasticSearch:
    a. Decrease the number of ElasticSearch shards for the index
27. A company collects IoT sensor data from devices on its factory floor for a year and stores the data in Redshift for daily analysis. It has been determined that at an expected ingestion rate of 2 TB per day, the cluster will be undersized in less than 4 months. A long term solution is needed, where most queries only reference the most recent 13 months of data but there are also quarterly reports that need to query all the data generated from the past 7 years. The CTO is concerned with costs, admin efforts, performance of a long term solutions:
    a. Create a daily job in Glue to unload records older than 13 months to S3 and delete those records from Redshift
    b. Create an external table in Redshift tot point to the S3 location
    c. Use Redshift Spectrum to join data that is older than 13 months
28. If migration takes more than week to transfer over the network:
    a. Use Snowball devices for offline migration
29. Company has raw data in JSON sent without a predefined schedule through a Kinesis Firehose delivery stream to an S3 bucket. A Glue crawler is scheduled to run every 8 hours to update the schema in the data catalog of the tables stored in the S3 bucket. The data is analyzed using Apache Spark SQL on EMR to set up with Glue Data Catalog as the metastore. Occasionally, the data received is stale and most up to date data needs to be provided:
    a. Run Glue crawler from a Lambda triggered by an S3:ObjectCreated:* event notification on S3 bucket
30. A company developed an Apache HIVE script to batch process data stored in S3. The script needs to run once everyday and store the output to S3. The company tested the script and it runs within 30 minutes on a small local 3 node cluster. The most cost effective solution for schedule and executing the script:
    a. Create a Lambda to spin up an EMR cluster with a Hive execution step
    b. Set KeepJobFlowAliveWhenNoSteps = false and disable the termination protection flag
    c. Use CW events to schedule the Lambda to run daily
31. Company launches its global website: all transaction data is stored in RDS and curated historical transaction data is stod in Redshift in the us-east-1 region. The BI team wants to provide a dashboard via QS. During development, a team in Japan provisioned QS in ap-northeast-1 and had difficulty connecting QuickSight from ap-northeast-1 to Redshift in us-east-1:
    a. Create a new SG for Redshift in us-east-1 with an inbound rule authorizing access from the appropriate IP address range for the QS servers in ap-northeast-1
32. A company with millions of users collect data on an hourly basis and store it in an S3 data lake. The company runs analyses on the last 24 hours of data flow logs for abnormality detection and to troubleshoot / resolve user issues. The company analyzes historical logs dating back to 2 years to discover patterns and look for improvement opportunities. The data flow logs contain many metrics and there are 10 billion events every day:
    a. For optimal performance: in Apache ORC, partition by date and sort by source IP

33. A company is using a Redshift cluster with dense storage nodes to store sensitive data. An audit found that the cluster is unencrypted. Compliance requirements state that a database with sensitive data must be encrypted through a hardware security module (HSM) with automated key rotation. To achieve compliance, need:
    a. Set up a trust connection with HSM using a client and server certificate with automatic key rotation
    b. Create a new HSM encrypted Redshift cluster and migrate the data to the new cluster
34. A company is doing a POC for an ML project using SageMajer with a subset of existing on premise data hosted in a 3 TB data warehouse. For the POC, Direct Connect is established and tested. To prepare the data for ML, analysts are performing data curation and want to perform this in multiple steps. The company needs the faster solution to curate the data for this project:
    a. Ingest data into S3 using DMS, use Glue to perform data curation and store the data in S3 for ML processing
35. A company has CSV data stored in S3 within a Glue DAta Catalog and this data should be joined with data from a call center stored in Redshift as part of a daily batch process. The Redshift cluster is already under a heavy load so the solution must be managed, serverless, well functioning and minimize the load on the existing Redshift cluster. The solution should also require minimal effort and development activity:
    a. Create an external table using Redshift Spectrum for call center data and perform a join with Redshift
36. An analyst using QuickSight for data visualization across multiple datasets generated by apps. Each app stores files within a separate S3 bucket. Glue DAta Catalog is used as a central catalog across all app data in S3. a new app stores its data within a separate S3 bucket. After updating the catalog to include the new app data source, the analyst created a new QS data source from Athena but the import into SPICE failed:
    a. Edit the permissions for the new S3 bucket from within the QuickSight console
37. Analyze market trend data that comes from 5 different data sources in large volume. Want to utilize Kinesis. SQL like queries are used to analyze trends and want to send notifications based on significant patterns in the trends. Want to save the data to S3 for archival and historical preprocessing and use AWS management services wherever possible. Want to implement the lowest cost solution:
    a. Publish data to a Kinesis Data Stream and deploy a Kinesis Data Analytics stream for analyzing trends
    b. Configure a Lambda as an output to send notifications using SNS
    c. Configure Kinesis Firehose on the Kinesis Data Stream to persist data to S3 bucket
38. A company uses Athena to query its global datasets. The regional data is stored in S3 in us-east-1 and us-west-2 regions and it is not encrypted. To simplify the query process and manage it centrally, the company wants to use Athena in us-west-1 to query data from S3 in both regions. The solution should be as low cost as possible
    a. Run Glue crawler in us-west-1 to catalog datasets in all regions and run Athena queries in us-west-2
39. A company is building a data warehouse solution on Redshift. The company is loading hundreds of files into the fact table created in the Redshift cluster. Want the solution to achieve the highest throughput and optimally use cluster resources when loading data into the table. To meet these requirements:
    a. Use a single COPY command to load the data into the Redshift cluster
40. An analyst is defining a solution to interactively query datasets with SQL using a JDBC connection. Users join data stored in S3 in Apache ORC format with data stored in OpenSearch and Aurora MySQL:
    a. For most up to date results, query all the datasets in place with Apache Presto running in EMR
41. A company has a streaming app that needs to collect, analyze data to provide near real time feedback on issues within 30 seconds. The company requires a consumer app to identify issues. The data will be streamed in JSON format and schema can change over time:
    a. Send the data to Kinesis Data Streams and configure a Kinesis Data Analytics for SQL app for Apache Flink as the consumer app to process and analyze the data
42. Need to ETL streaming data from web logs as it is streamed in for analysis in Athena. The ETL does not strictly need to happen in real time but transforming the data within a minute is desirable. A viable solution to this requirement:
    a. Perform any initial ETL using Kinesis and store the data to S3
    b. Trigger a Glue ETL to complete the transformations needed
43. A company receives files from external parties in EC2 throughout the day. At the end of the day, the files are combined into a single file, comparedded into GZIP and uploaded to S3. The total size of all the fie;s is close to 100 GB daily and once the files are uploaded to S3, a Batch program executes a COPY command to load the field into a Redshift cluster. To accelerate the copy process:
    a. Split the number of files so they are equal to a multiple of the number of slices in the Redshift cluster
    b. GZIP and upload the files to S3 and run the COPY command on the files
44. A company has thousands of drivers serving millions of unique customers everyday. The company migrates an existing data mart to Redshift and the existing schema includes the following tables: a trip table for information on completed rides, a drivers table for driver profiles, a customer table holding customer profile information. The company analyzes trip details by date and destination to examine profitability by region. The drivers data rarely changes and the customers data frequently changes. To provide optimal query performance:
    a. Use distribution style KEY (default) for the trips table and sort by date
    b. Use distribution style ALL for the drivers table (rarely changes)
    c. Use distribution style EVEN for the customers table (frequently changes)

45. Teams use Apache HIVE on the EMR cluster with EMRFS to query data stored within each team's S3 bucket. The EMR cluster has Kerberos enabled and is configured to authenticate users from the corporate Active Directory. The data is highly sensitive so access must be limited to the members of each team:
    a. For the EMR clusters EC2 instances, create a service role that grants no access to S3
    b. Create 3 additional IAM roles, each granting access to each teams specific bucket
    c. Add the service role of the EMR clusters EC2 instances to the trust policies for the additional IAM roles
    d. Create a security config mapping for the additional IAM roles to Active Directory user groups for each team
46. A company wants to create a data lake in S3. The company wants to create tiered storage based on access patterns and cost objectives. The solution must include support for JDBC connections from legacy clients, metadata management that allows federation for access control and batch based ETL using PySpark and Scala. Operational management should be limited. To meet these requirements:
    a. Glue Data Catalog for metadata management and Glue for Scala based ETL
    b. Athena for querying data in S3 using JDBC drivers
47. Company wants to optimize the cost of its DA platform: ingesting CSV/JSON files in S3 from various data sources. Incoming data is expected to be 50 GB per day. The company is using Athena to query the raw data in S3 directly. Most queries aggregate data from the past 12 months and data older than 5 years is infrequently queried. The typical query scans about 500 MB of data and is expected to return results in less than 1 minute. The raw data must be retained indefinitely for compliance requirements:
    a. Use Glue ETL job to compress, partition, convert the data into a columnar data format
    b. Use Athena to query the processed dataset
    c. Configure a lifecycle policy to move the processed data into S3 Standard IA class 5 years after object creation
    d. Configure a lifecycle policy to the raw data into S3 Glacier for long term archival 7 days after object creation
48. A company collects data in real time from sensors and wants to receive notifications when bad data is detected within 10 minutes, which must be delivered ASAP. System must be highly available: autoscaling solution that scales when monitoring feature is implemented in other views and for notification system to be subscribed to SNS:
    a. Create an Amazon Managed Streaming for Apache Kafka cluster to ingest the data
    b. Usee an Apache Spark Streaming with Apache Kafka consumer API in an automatically scaled EMR cluster to prices the incoming data
    c. Use the Spark Streaming app to detect the known event sequence and sent to SNS
49. Data should be delivered and managed by AWS in ner real time to ElasticSearch:
    a. Use Kinesis Firehose
50. A company stores data in RDS and wants a solution to store, analyze historical data. The most recent 6 months of data will be queried frequently for analytics workloads. This data is several TBs large. Once a month, historical data for the last 5 years must be accessible and will be joined with the more recent data. To optimize performance and cost:
    a. Incrementally copy data from RDS to S3. Load and store the more recent 6 months of data in Redshift
    b. Configure a RedShift Spectrum table to connect to all historical data
51. A company uses S3 as its data lake and sets up a data warehouse using a multi-node Redshift cluster. The data fields in the data lake are organized in folders based on the data source of each data file. All data files are loaded to one table in the Redshift cluster using a separate COPY command for each data file location. With this approach, loading all the data files in Redshift takes a long time to complete. Users want a faster solution with little or no increase in cost while maintaining the segregation of the data files in the S3 data lake:
    a. Use EMR to copy all the data files into 1 folder and issue a COPY command to load data into Redshift
52. Company uses Athena for ad-hoc queries on data in S3: wants to implement additional control to separate query execution / query history among users running in the same AWS account to comply with internal security policies:
    a. Create an Athena workgroup for each given use case, apply tags to the workgroup and create an IAM policy using the tags to apply appropriate permissions to the workgroup
53. A company wants to use an automatic ML random cut forest algorithm to visualize complex real world scenarios. The team working on this project is non technical and is looking for an out of the box solution that will require the least amount of management overhead:
    a. Use QuickSight to visualize the data and use ML powered forecasting to forecast the key business metric
54. A DA team recently created multiple product dashboards for the average price per product using QuickSight. The dashboard was created from CSV files in S3. The team wants to share the dashboards with external product owners by creating individual users in QS. For compliance reasons, restricting access is a key requirement. The product owners should view only their respective analysis in the dashboard reports. To allow product owners to view only their products in the dashboard:
    a. Create dataset rules with row level security
55. A company wants to improve the data load time of a dashboard. Data is collected as CSV files and stored in S3 bucket partitioned by date. The data is loaded to Redshift data warehouse for frequent analysis. The data volume is up to 500 GB per day. To improve the data loading performance:
    a. Split the large CSV files and use a COPY command to load data into Redshift

56. A company has a 500 TB data warehouse in Redshift. New data is imported every few hours and read-only queries are run throughout the day and evening. There is a heavy load with no writes for several hours each morning on business days. During those hours, some queries are queued so it takes a long time to execute. To optimize query execution and avoid any downtime, the most cost effective solutions:
    a. Enable concurrency scaling in the workload management queue
57. A company analyzes its data in a Redshift data warehouse, which is a cluster of 3 dense storage nodes. Due to a recent acquisition, the company needs to load an additional 4 TB of user data into Redshift. The engineering team will combine all the user data and apply complex calculations that require I/O intensive resources. The company needs to adjust the clusters capacity to support the change in analytical and storage requirements:
    a. Resize the c;suter using Elastic Resize with dense compute nodes
58. A company stowed data that includes PII in S3. The company allows its analysts to launch their own EMR cluster and run analytics reports within the data. To meet compliance requirements, the company must ensure the data is not publicly accessible throughout this process and must ensure the individual EMr clusters created by the analysts are not exposed to the public internet. To meet this compliance requirement with least amount of effort:
    a. Enable the block public access setting for EMR at the account level before any EMR cluster is created
59. A company created a dashboard that visualizes and analyzes time sensitive data. The data will come in through Kinesis Data Firehose with buffer interval set to 60 seconds. The dashboard must support near real time data:
    a. Use OpenSearch as the endpoint for Firehose.
    b. Set up an OpenSearch dashboard (Kibana) using data in OpenSearch with desired analysis / visualizations
60. A team wants to identify a high performing long term storage service for their data based on the following requirements: the data size is approximately 32 TB uncompressed, there is a low volume of single row inserts each day, there is a high volume of aggregation queries each day, multiple complex joins are performed and the queries typically involve a small subset of the columns in a table. The most performant solution is:
    a. Redshift
61. A company uses Apache Hive on EMR for ad-hoc queries. Users are complaining of slow performance. An analyst notes that about 90% of queries are submitted 1 hour after the market opens and HDFS utilization never exceeds 10%. To help address the performance issues:
    a. Create instance group configurations for core and task nodes
    b. Create an automatic scaling policy to scale out the instance groups based on YARM Memory Available Percentage metric
    c. Create an automatic saling policy to scale in the instance groups based on the CloudWatch YARN Memory Available Percentage metric
62. A company has been performing analytics on log data generated by its apps. There has been a recent increase in the number of concurrency analytics jobs running, and the overall performance of existing jobs is decreasing as the number of new jobs is increasing. The partitioned data is stored in S3 One Zone IA and the analytic processing is performed on EMR clusters using the EMRFS with consistency view enabled. An analyst determined that it is taking longer for the EMR task nodes to list objects in S3. The most likely action to increase the performance of accessing log data in S3:
    a. Increase the RCUs for the shared DynamoDB table
63. Management wants a dashboard to monitor current revenue against their annual revenue goal:
    a. Use KPIs on QuickSight
64. A company developed Glue jobs to validate / transform its data from S3 and load it into RDS for MySQL in batches once everyday. The ETL jobs read the S3 data using a DynamicFrame. Currently, the ETL developers are experiencing challenges in processing only the incremental data on every run, as the Glue job processes all the S3 input data on each run. To solve the issue with minimal coding effort:
    a. Enable job bookmarks on the Glue jobs
65. A company has a service for accepting payments that uses the DynamoDB Encryption Client with KMS managed keys to encrypt the sensitive data before writing the data to DynamoDB. The finance team should be able to load this data into Redshift and aggregate the va;ues within the sensitive fields. The Redshift cluster is shared with other analysts from different business units. To accomplish this task efficiently and securely:
    a. Create a Lambda to process the DynamoDB stream and decrypt the sensitive data using the same KMS key
    b. Save the output to a restricted S3 bucket for the finance team
    c. Create a finance table in Redshift that is accessible to the finance team only
    d. Use COPY command to load the data from S3 to the finance table
66. Building a data lake and need to ingest data from a relational DB that has time series data. The company wants to use managed services to accomplish this. The process needs to be scheduled daily and bring incremental data only from the source into S3. The most cost effective approach to meet this requirement is:
    a. Use Glue to connect to the data source using JDBC drivers
    b. Ingest incremental records only using job bookmarks
67. An ecommerce website uses an on-premise PostgreSQL database as its main OLTP database. Want to perform analytical queries on it but the architect recommended against doing it off of the main database:
    a. Use DMS to replicate the database to RDS

68. A Redshift DB contains sensitive user data. Logging is necessary to meet compliance requirements. The logs must contain database authentication attempts, connections, disonnections. The logs must also contain each query run against the database and record which database user ran each query. To create the required logs:
    a. Enable audit logging for Redshift using AWS console or CLI
69. Once a month, a company receives a 100 MB CSV file compressed with GZIP. The file contains 50,000 listing records and is stored in Glacier. A data analyst needs to query a subset of the data for a specific vendor:
    a. Load the data to S3 and query it with S3 Select
70. A company that monitors weather conditions from sites is setting up a solution to collect data from station A (10 sensors) and station B (5 sensors). Each sensor has a unique ID and data collected using Kinesis Data Streams. Based on total incoming and outgoing data throughput, a single Kinesis Data Stream with 2 shards is created. Two partition keys are created based on the station names. During testing, there is a bottleneck on data coming from station A but not from station B. It is confirmed that the total stream throughput is still less than the allocated Kinesis Data Streams throughput. To resolve the bottleneck without increasing the overall cost and complexity of the solution, while retaining data collection quality requirements:
    a. Modify the partition key to use the sensor ID instead of the station name
71. A company developed a new website that uses Kinesis Firehose to deliver full logs from AWS WAF to an S# bucket. The company is now seeking a low cost option to perform this infrequent analysis with visualizations of logs in a way that requires minimal development effort. To meet these requirements:
    a. Use Glue crawler to create and update a table in the Glue Data CAtalog from the logs
    b. Use Athena to perform ad-hoc analysis and use QuickSight for data visualizations
72. A company has a central data lake to run analytics across different departments, where each uses a separate AWS account and stores its data in a S3 bucket in that account. Each AWS account uses the Glue Data Catalog as its data catalog. There are different data lake access requirements based on roles and analysts should only have read access to their departmental data. Senior analysts can have access to multiple departments including theirs, but for a subset of columns only. To achieve the required access patterns to minimize costs and tasks:
    a. Set up an individual AWS account for the central sata lake
    b. Use LakeFormation to catalog the cross account locations
    c. On each individual S3 bucket, modify bucket policy to grant S3 permissions to the Lake Formation service linked role
    d. Use Lake Formation permissions to add a fine grained access controls to allow specific analytics to view specific tables and columns
73. A company wants to improve satisfaction for its system by adding more features to its engine. Each sensor only pushes its nested JSON data into Kinesis Data Streams using KPL in Java. Status from a set of failed sensors showed that, when a sensor is malfunctioning, its recorded data is not always sent to the cloud. The company needs a solution that offers near real time analysis on the data for the most updated sensors:
    a. Update the sensor code to use the PutRecord(s) call from the Kinesis Data Streams API with the Java SDK
    b. Use Kinesis Data Analytics to enrich the data based on the companies anomaly detection SQL script
    c. Direct the output of the app to Kinesis Firehose delivery stream and enable the data transformation feature to flatten the JSON file
    d. Set the Kinesis Firehose destination to an OpenSearch cluster
74. A company has different suborgs and each sub org sells its products / services in various countries. The leadership wants to quickly identify which cyborg is the faster performer in each country. All sales data is stored in S3 in Parquet format. To provide the request with the least amount of effort:
    a. Use QuickSight with Athena as the data source and use what maps as the visual type
75. A company has 1 million scanned documents stored as image files in S3. The documents contain typewritten forms with PII. The company has developed a ML algorithm to extract the metadata values from the scanned documents. The company wants to allow internal analysts to analyze and find apps using the PII. Cost control is secondary to query performance. To organize the images and metadata to drive insights:
    a. Index the metadata and the S3 location of the image file in OpenSearch Service.
    b. Allow the data analysts to use OpenSearch dashboards (Kibana) to submit queries to OpenSearch cluster
76. Want to capture data from its app and make the data available for analysis immediately. The data record size will be approximately 20 KB. The company is concerned about achieving optimal throughput from each device. Additionally, the company wants to develop a data stream processing app with dedicated throughput for each consumer. To achieve this goal:
    a. Have the app call PutRecords API to send data to Kinesis Data Streams
    b. Use Enhanced Fan Out feature while consuming the data
77. The most common identity principles for authentication supported by IoT for mobile applications:
    a. Cognito Identities
78. An organization has hundreds of TBs of data stored within its on premise data centers and data is being produced at the rate of GBs per second and could be consumed within 3 days. As part of their AWS cloud migration:
    a. Transfer historical data using Snowball and use Kinesis Data Streams for ongoing data collection

79. A company wants to improve its BI capabilities given that: the operations team reports are run hourly for the current month's data, the sales team wants to use multiple QS dashboards to show a rolling view of the last 30 days based on several categories and wants to view data as soon as it reaches the reporting backend, the finance team reports are run daily for the last months data and once a month for the last 24 months data. Currently, there is 400 TB of data in the system with an expected additional 100 TB added every month. The company wants a cost effective solution ASAP:
    a. Store the last 2 months of data in Redshift and the rest of the months in S3
    b. Set up an external schema and table for Redshift Spectrum
    c. Configure QuickSight with Redshift as the source
80. A company wants to perform ML and analytics on the data residing in its S3 data lake. There are 2 data transformation requirements that will enable the consumers within the company to create reports: daily transformations of 300 GB of data with the different file formats landing in S3 at a scheduled time and one time transformations of TBs of archived data residing in the S3 data lake. To meet these requirements:
    a. For daily incoming data, use Glue crawlers to scan and identity the schema
    b. Use Glue Workflows with Glue jobs to perform transformations
    c. For archived data, use EMR to perform data transformations
81. A company uses sensors to collect data and wants a near real time solution that can ingest the data securely at scale. The solution should be able to remove the patient's PII from the streaming data and store the data in durable storage. To meet these requirements with the least operational overhead:
    a. Ingest the data using Kinesis Firehose to write the data to S3
    b. Implement a transformation Lambda that parses sensor data to remove all PII
82. A company is migrating its existing on premise ETL jobs to EMR. The code consists of a series of jobs in Java. The company needs to reduce overhead for the sysadmins without changing the underlying code. Due to the sensitivity of the data, compliance requires that the company uses root device volume encryption on all nodes in the cluster. Corporate standards require that environments by provisioned through CloudFormation when possible:
    a. Create a custom AMI with encrypted root device volumes
    b. Configure EMR to use the custom AMI using Custom Ami Id property in the CloudFormation template
83. A company uses IoT sensors attached to trucks to collect data for its global delivery fleet. The company currently sends the data in small CSV files to S3. The files are then loaded into a 10 node Redshift cluster with 2 slices per node and queries using Athena and Redshift. The company wants to optimize the files to reduce the cost of querying and also improve the speed of data loading into the Redshift cluster:
    a. Use Glue to convert the files from CSV to Apache Parquet to create 20 Parquet files
    b. Copy the files into Redshift and query te files with Athena from S3
84. A company with millions of users around the globe wants to improve its ecommerce analytics capabilities. Data is uploaded directly to S3 as compressed files. Several times each day, an app running on EC2 processes the data and makes search options and reports available for visualization. The company wants to make website clicks and aggregated data available to editors in minutes to enable them to connect with users effectively.
    a. Use Kinesis Firehose to upload compressed and batched clickstream records to OpenSearch service
    b. Use OpenSearch dashboards (Kibana) to aggregate, filter, visualize the data stored in OpenSearch service to refresh content dashboards in real time
85. A company is streaming its high volume billing data (100 Mbps) to Kinesis Data Streams. An analyst partitioned the data in account_id to ensure that all records belonging to an account go to the same Kinesis shard and order is maintained. While building a custom consumer using Kinesis Java SDK, the analyst notices that sometimes the messages arrive out of order for account_id. Additionally, it is discovered that the messages are out of order seem to be arriving from different shards for the same account)id and are seen when a stream resize runs
    a. Consumer is not processing the parent shard completely before processing the child shards after steam resize
    b. The analyst should process the parent shard completely first before processing the child shards
86. A company consumes a stream of posts that are sent to a Kinesis Data Stream partitioned on user_id. A Lambda retrieves the records and validates the content before loading the posts into an OpenStack cluster. The validation process needs to receive the posts for a given user in the order they were received by the Kinesis Data Stream. During peak hours, the posts take more than 1 hour to appear in the OpenSearch cluster. A data analytics specialist must implement a solution that reduces this latency with the last possible operational overhead:
    a. Increase the number of shards in Kinesis Data Stream
87. A company launched a service that produces millions of messages everyday and uses Kinesis Data Streams as the streaming service. The company uses the Kinesis SDK to write data to Kinesis Streams. A few months after launch, an anaust found that write performance is significantly reduced. The analyst investigated the metrics and determined that Kinesis is throttling the write requests. The analyst wants to address this issue without significant changes to the architecture. To resolve:
    a. Increase the number of shards in the stream using UpdateShardCount API
    b. Choose partition keys in a way that results in a uniform record distribution across shards

88. A company must efficiently ingest and process messages from various connected devices / sensors. Most of the messages consist of a large number of small files that are ingested using Kinesis Data Streams then sent to S3 using a Kinesis Stream consumer app. The S3 message data is passed through a processing pipeline built in EMR running PySpark jobs. The data platform team manages data processing and is concerned about the efficiency and cost of downstream data processing. They want to use PySpark. To improve the efficiency of the data processing jobs and be well architected:
    a. Send the sensor and device data directly to a Kinesis Firehose delivery stream to send data to S3 with Parquet record format conversion enabled
    b. Use EMR running PySpark to process data in S3
89. A company is running its ETL process, part of which is to move data from S3 to the Redshift cluster. The company wants to use the most cost efficient method to load the dataset into Redshift:
    a. Use the COPY command with the manifest file to load data into Redshift and use temporary staging tables during loading
90. A university intends to use Kinesis Firehose to collect JSON formatted batches to S3. The data is captured from 50 scattered sensors and students will query the stored data using Athena to observe changes in a captured metric over time. Interest in the study has grown so reconsidering how data is stored:
    a. Partition the data by year, month and day then store in parquet format using Snappy compression
91. A company uses AWS DA tools to collect, ingest, store EHR data about its patients. The raw EHR data is stored in S3 in JSON format partitioned by hours, days, year and is updated every hour. The company wants us to maintain the data catalog and metadata in a Glue DAta Catalog to be able to access the data using Athena or Redshift Spectrum for analytics. When defining tables in the Data Catalog, the company has the following requirements: choose catalog table name and don't reply on catalog table naming algorithm, jeep the table updated with partition loaded in respective S3 bucket prefixes:
    a. Use Glue API CreateTable operation to create a table in Data Catalog
    b. Create Glue crawler and specify table as the source
92. A university has adopted a strategic goal of increasing diversity: the analytics team is creating a dashboard with visualizations to enable stakeholders to view historical trends. All access must be authenticated using Microsoft Active Directory. All data in transit and at rest must be encrypted:
    a. QS Enterprise Edition configured to perform identity federation using SAML2 and default encryption settings
93. An airline has been collecting metrics on flights for analysts and finished a POC to demonstrate insights to analysts to improve on time departures. The POC used objects in S3, which contained the metrics in CSV format and used Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance. To improve performance as the data lake grows:
    a. Compress the objects to reduce the data transfer I/O and use S3 bucket in the same region as Athena
    b. Preprocess CSV to Parquet to reduce I/O by fetching only the data blocks needed for predicates
94. A company uses EMR clusters for its workloads and manually installed third party libraries on the clusters by logging into the master nodes. To create an automated solution to replace the manual process:
    a. Place the required installation scripts in S3 and execute them using custom bootstrap actions
    b. Launch EC2 instance with Linux and install the required third party libraries on the instance
    c. Create an AMI and use that AMI to create the EMR cluster
95. Analyst needs to ensure that queries run in Athena cannot scan more than a prescribed amount of data for cost control purposes. Queries that exceed the prescribed threshold must be canceled immediately. To achieve
    a. For each workgroup, se the control limit for each query to the prescribed threshold
96. A company wants to collect large volumes of transactional data using Kinsi Data Streams for real time analytic: uses PutRecord to send data to Kinesis and wants to obtain exactly once semantics for the entire pipeline. To obtain:
    a. Design the app so it can remove duplicates during processing by embedding a unique ID in each record
97. A company uses Redshift as its data warehouse: a new table has columns that contain sensitive data. The data in the table will be referenced by several existing queries that run many times a day. An analyst needs to load 100 billion rows of data into the new table. Before doing so, the analyst must ensure that only members of the auditing group can read the column containing sensitive data. To meet requirements with low maintenance:
    a. Load all the data into the new table and grant the auditing group permission to read from the table
    b. Use the GRANT command in SQL to allow read only access to a subset of columns to the appropriate users
98. Company wants to enrich app logs in near real time and use the enriched dataset for further analysis: the app is running on EC2 instances across multiple AZs and storing its logs using CW logs. The enrichment source is stored in a DynamoDB table. To meet these requirements for event collection and enrichment:
    a. Use CW logs subscription to send data to Firehose and use Lambda to transform the data in Firehose delivery stream by enriching it with data in the DynamoDB table
    b. Configure S3 as the Firehose delivery destination
99. A gaming company stores each game's data in DynamoDB tables. To provide game search functionality to your users, you need to move that data over to ElasticSearch. To do so efficiently and as close to real time as possible:
    a. Enable DynamoDB streams and write a Lambda function

100. Company uses Kinesis SDK to write data to Kinesis Data Streams: compliance requirements state data must be encrypted at rest using a key that can be rotated. The company wants to meet those encryption requirements with minimal coding effort:
    a. Create a customer master key (CMK) in KMS and assign the CMK as an alias
    b. Use the encryption SDK and provide it with the key alias to encrypt and decrypt the data
101. A team within a shared workspace company wants to build a centralized logging system for all web logs generated by the space reservation system. The company has a fleet of EC2 instances that process requests for shared space reservations on its website. The team want sot ingest all web logs into a service that provides a near real time search engine but does not want to manage maintenance or operation of the logging system:
    a. Set up a CW agent to stream web logs to CW logs and subscribe the KinesisFirehose delivery stream to CW
    b. Choose OpenSearch as the end destination of the web logs
102. A company wants to research the last 3 months of user activities: it has millions of users and 1.5 TB of uncompressed data is generated each day. A 30 node Redshift cluster with 2.56 TB of SSD storage for each node is required to meet the query performance goals. Want to run an additional analysis on a year's worth of historical data to examine trends for which features are most popular. Analysis is done once a week:
    a. Keep data from the last 90 days in Redshift
    b. Move data older than 90 days to S3 and store it in Parquet format partitioned by date
    c. Use Redshift Spectrum for additional analysis
103. A bank operates in a regulated environment: compliance requirements for the bank's country says that customer data for each state should only be accessible by bank employees located in the same state. Bank employees in one state should not be able to access data for customers who provided an address in a different state. The banks marketing team hired an analyst to gather insights from the customers data for a new campaign being launched in certain states. Currently, data linking each customer account to its home state is stored in CSV files within a single folder in a private S3 bucket. The total size of the S# folder is 2GB uncompressed. Due to the country’s compliance requirements, the marketing team is not able to access the folder. The analyst is responsible for ensuring that the marketing team gets a one time access to customer data for their campaign analytics project, while being subject to all the compliance requirements and controls:
    a. Load tabular CSV data from S3 to QuickSight Enterprise edition yb directly importing it as a data source
    b. Use the built in row level security feature in QuickSight to provide marketing employees with appropriate data access under compliance controls
    c. Data QuickSight data sources after project is complete
104. A company uses Kinesis Analytics SQL app with Kinesis Data Stream as its source. The source sends three non null fields to the app: col_1, col_2, col_3. An analyst has a CSV file that maps a small number of col_3 values to code. The analyst needs to include the cost (if it exists) as an additional output of the Kinesis Analytics app. To do so while minimizing costs:
    a. Store the mapping file in S3 and configure it as reference data source for the Kinesis Analytics app
    b. Change the SQL query in the app to include a join to the reference table and add the code field to select columns
105. A company uses DynamoDB as the database for user support apps. The company is developing a new version of the app that stores a PDF file for each case ranging in size from 1 to 10 MB. the files should be retrievable whenever the case is accessed in the application:
    a. Store the file in S3 and object key as an attribute in DynamoDB table
106. A website captures user activity and sends data to Kinesis Data Streams. The company wants to design a cost effective solution to process the data to create a timeline of user activity within a session. The solution must be able to scale depending on the number of active sessions:
    a. Include a session ID in the data from the website and use it as a partition key for the stream to perform the processing
        i. Partition by session ID will allow a single processor to process all the actions for a user session in order
    b. Deploy the consumer app on EC2 instances an in auto saling group
    c. Use Lambda to reshard the stream based on CW alarms
        i. Lambda calls the UpdaeShardCount API action to change the number of shares in the stream: KCL will automatically manage the number of processors to match the number of shards
107. A company ingesta large set of clickstream data in nested JSON format from different sources and stores it in S3. Analysts need to analyze this data in combination with data stored in a Redshift cluster. Analysts want to build a cot effective and automated solution for this need:
    a. Use the Relationalize class in Glue ETL job to transform the data and write the data back to S3
    b. Use Redshift Spectrum to create external tables and join with the internal tables
108. A company has an app that ingests streaming data: need to analyze this stream over a 5 minute time frame to evaluate the estream for anomalies with random cut forest and summarize the current count of status code. The source and summarized data should be persisted for future use. To do so while keeping costs low:
    a. Ingest the data stream using Kinesis Data Streams and have a Kinesis Analytics app evaluate the stream over a 5 minute window using random cut forest function and summarize the count of status codes
    b. Persist the source and results to S3 through output delivery to Kinesis Firehose

109. A company collected more than 100 TB of log files in the last 2 years. The files are stored as raw text in an S3 bucket. Each object has a key of form year-month-day-_log_hhmmss which is the time the log was created. A table was created in Athan that points to the S3 bucket. One time queries are run against a subset of columns in the table several times an hour. An analyst must make changes to reduce the cost of running these queries and management wants a solution with minimal maintenance overhead:
    a. Add a key prefix date=year-month-day/ to S3 objects to partition the data
    b. Convert log field to Parquet format
    c. Drop and recreate the table with PARTITIONED BY clause
    d. Run the MSCK REPAIR TABLE statement
110. A company needs to implement a near real time fraud prevention feature for its ecommerce site. User and order details need to be delivered to a SageMaker endpoint to flag suspected fraud. The amount of input data needed for the inference could be as much as 1.5 MB. for the lowest overall latency:
    a. Create an Amazon Managed Streaming for Kafka Cluster and ingest the data for each order into a topic
    b. Use a Kafka consumer running on EC2 instances to read these messages and invoke the SageMaker Endpoint
111. A company is migrating its on-premises legacy Hadoop cluster with its associated data processing scripts and workflow to an EMR environment running the latest HAdiio release. The developers want to resume the Java code that was written for data processing jobs for the on premises clusters:
    a. Compile the JAva program for the desired HAdoop version and run it using a custom_jar step on EMR cluster
112. A company wants to perform analytics on data in large S3 objects using EMR. A Spark job repeated queries the same data to populate an analytics dashboard. The analytics team wants to minimize the time to load the data and create the dashboard. To improve the performance:
    a. Load the data into Spark DataFrame and use S3 Select to retrieve the data necessary for the dashboards form S3 objects
113. An engineer needs to create a dashboard to display social media trends during the last hour of a large company event. The dashboard needs to display the associated metrics with a latency of less than 1 minute.
    a. Publish the raw social media data to a Kinesis Firehose delivery stream
    b. Use Kinesis Analytics for SQL apps to perform a sliding window analysis to compute the metrics with a latency of less than 1 minute
    c. Configure a Lambda to save the stream data to a DynamoDB table
    d. Deploy a real time dashboard hosted in an S3 bucket to read and display the metrics data stored in DynamoDB table
114. A company is receiving new property listing data from its agents through CSV files everyday and storing these files in S3. The analytics team created a QuickSight report that uses a dataset imported from the S3 files. The team wants the visualization report to reflect the current data up to the previous day:
    a. Schedule the dataset to refresh daily
115. Looking to reduce latency down from your big data processing job that operates in Singapore but the source data is in virginia. The big data job must always operate against the latest version of the data:
    a. Enable S3 Cross Region replication
116. A company is providing analytics services to its HR departments (only access data through their /b/i tools) which run Presto queries on EMR clusters that use EMRFS. The marketing analyst must be granted access to the advertising table only. The HR analyst must be granted access to personnel table only
    a. Create separate IAM roles for the marketing and HR users
    b. Assign the roles with Glue resource based policies to access their corresponding tables in Glue Data Catalog
    c. Configure PResto to use the Glue Data Catalog as the Apache Hive metastore
117. A company uses EMR for its analytics workloads. During the company's annual security audit, the security team determined that none of the EMRs clusters root columns are encrypted. The security team recommends the company encrypt its EMR clusters root volume ASAP. To meet these requirements:
    a. Specify local disk encryption in a security configuration
    b. Recreate the cluster using the newly created security configuration
118. Daily Spark jobs are run against files created by a Kinesis Firehose Pipeline in S3. Due to low throughput, each of the files created by Kinesis Firehose is about 100 KB. to optimize the Spark job as best as possible to query the data efficiently:
    a. Consolidate the files on a daily basis using Data Pipeline
119. A team developed a Spark Streaming app that performs real time transformations on an on-premise Apache Kafka cluster and delivers the data in real time to S3. As part of a cloud migration and switch to Kinesis for streaming store:
    a. Produce data using Spark streaming and read data with Spark streaming from Kinesis data Streams to write to S3
120. An application on AWS generates frequent logs on S3. The team needs to analyze the logs to understand various patterns of the application failures and come up with a plan to fix those issues. The most cost effective option that requests the last engineering effort would be:
    a. Create Glue Data Catalog metadata and analyze the logs via Athena

121. Storing gaming data for a game that is becoming increasingly popular. An average game data contains 80 KB of data. You expect 400 games to be written per second to your database. Additionally, a lot of people would like to receive this game data and you expect about 1800 eventually consistent reads per second. To provision the DynamoDB table:
    a. 1 WCU = 1 KB per second so need 80 KB * 400 per second = 32000 WCU
    b. 1 RCU = 2 eventually consistent reads per second of 4 KB so need 1800 * 80 / 8 = 18000 RCU
122. An enterprise wants to leverage Redshift to query data. The data is produced at the rate of 5 PB of historical data and another 3 TB per month. The non historical, ongoing data should be available with less than 2 days delay. You are tasked with finding the most efficient data transfer solution into S3. It is recommended to:
    a. Use Snowball to transfer the historical data
    b. Establish Direct Connect and do a daily upload pf newly created monthly data directly into S3
123. An Elasticsearch domain has been installed within a VPC. Two methods that could be employed to securely allow access to Kibana from outside the VPC:
    a. Set up a reverse proxy server between your browser and ElasticSearch service
    b. Set up an SSH tunnel with port forwarding to allow access on port 5601 (Kibana runs on port 5601 by default)
124. The analytics team at an e-commerce company uses Apache Hive on EMR. Several analysts have reported sub-par performance for the cluster during the morning peak load hours when 95% of the daily queries are executed by the analysts. The analytics team noted that HDFS usage never surprises 10%. To resolve these performance issues:
    a. Set up instance group configurations for core and task nodes
    b. Leverage the CW YARN Memory Available Percentage metric to configure automatic scaling policies to scale out and scale in the instance groups
125. A company uses two AWS accounts for accessing various AWS services. The analytics team has just configured a S3 bucket in Account A for writing data from the Redshift cluster provisioned in Account B. The team has noticed that the files created in the S3 bucket using UNLOAD command from the Redshift cluster are not accessible to the bucket owner user of Account A that created the S3 bucket. The reason for this denial of permission for resources belonging to the same Account:
    a. By default, an S3 object is owned by the AWS account that uploaded it so the S3 bucket owner will not imply likely have access to the objects written by the Redshift cluster
126. Processing data using a long running EMR cluster and want to ensure that it can recover data in case an entire AZ goes down, as well as process the data locally for the various Hive jobs you plan on running. To do this at a minimal cost:
    a. Store the data in S3 and keep a warm copy in HDFS
127. A healthcare analytics company wants to use QS to develop dashboards for analyzing health metrics. A team of 10 analysts will author these dashboards that will later be shared with 1000 health care professionals. The given health data is gathered from multiple research institutes and the data is later uploaded to S3 every 24 hours. The data is divided into years and months, and saved in Apache Parquet format. The company's primary data catalog is Glue DAta Catalog and the data querying is handled via Athena. At any point in time, the dashboards query from a total of 200 GB of uncompressed data. The most cost effective solution to address the given scenario:
    a. Set up QS Enterprise Edition: create 10 author users and 1000 reader users
    b. Configure an Athena data source and import the data into SPICE which is then automatically refreshed every 24 hours
128. A company wants to use AWS for its connected cab app that would collect sensor data from its electric cab fleet to give drivers dynamically updated map information. The company would like to build its new sensor service by leveraging fully serverless components that are provisioned and managed automatically by AWS. The development team does not want an option that requires the capacity to be manually provisioned, as it does not want to respond manually to changing volumes of sensor data. To develop this service:
    a. Ingest the sensor data in an SQS standard queue, which is polled by a Lambda in batches
    b. The data is written to an auto scaling DynamoDB for downstream processing
129. A research agency has deployed two autonomous underwater vehicles to track parameters. Vehicle A has 20 sensors whereas Vehicle B has 10 sensors. Each sensor is identified by a unique ID. Kinesis Data Streams is used to gather data from each sensor: a single stream with two shards is configured based on the total incoming and outgoing data throughput. Two partition keys are generated based on the names of the vehicles. During initial testing, data from Vehicle A experiences a bottleneck whereas data from Vehicle B does not. The overall stream throughput has been validated to be less than the assigned Kinesis Data Streams throughput. To address this bottleneck without increasing total cost and complexity:
    a. Change the partition key to use sensor ID instead of vehicle name
130. The analytics team is building an ElasticSearch based index for all the existing files in S3. to build this index ,it only needs to read the first 250 bytes of each object in S3 (which contains the metadata about the content of the file itself). There are over 100,000 files in your S3 bucket, adding up to 50 TB of data. To build this index the most efficiently:
    a. Create an application that will traverse the S3 bucket, issue a ByteRangeFetch for the first 250 bytes and store that information in ElasticSearch
    b. Create an application that will use the S3 Select ScanRange parameter to get the first 250 bytes and store that information in ElasticSearch

131. A retail company uses RDS to store sales data. For the analytics workloads that require high performance, only the last 6 months of data (approximately 50 TB) will be frequently queried. At the end of each month, the monthly sales data will be merged with the historical sales data for the last 5 years, which should also be available for analysis. The CTO at the company is looking at a cost optimal solution aht offers the best performance for this use case:
    a. Export RDS data to S3 and schedule a AWS Data Pipeline for incremental copy of RDS data to S3
    b. Load and store the last 6 months of data from S3 into Redshift
    c. Configure a Redshift Spectrum table to connect to all the historical data in S3
132. An e-commerce company stores all transaction data in RDS and the transformed transaction data in Redshift in the us-east-1 region. The analytics team wants to improve the user experience by developing a BI dashboard that highlights the sales trends over the last year. A team in India configured QS in ap-south-1 region during development. The team is experiencing connectivity issues between QS in ap-south-1 region and Redshift in us-west-1 region:
    a. Configure a new SG for Redshift in us-east–1 with an inbound rule authorizing access from the appropriate CIDR address block for the QS servers in ap-south-1
133. A company runs multiple gaming platforms that need to store game information. The company wants to move to AWS to scale reliably to millions of concurrent users and requests while ensuring consistently low latency measured in single digit milliseconds. The development team at the company is evaluating multiple in memory data stored with the ability to power its on demand, live leaderboard. The company's leaderboard required high availability, low latency and real time processing to deliver customizable user data for the community of its users.
    a. Develop the leaderboard using ElasticCache Redis as it meets the in memory, high availability, low latency requirements
    b. Develop the leaderboard using DynamoDB with DAX as it meets the in memory, high availability, low latency requirements
134. A trading firm collects daily stock trading data from exchanges and stores it in a data warehouse. The analytics team at the firm needs a solution that streams data directly into the data repository but should also allow SQL based data modifications when needed. The solution should facilitate complex analytical queries that execute in the fastest possible time. The solution should also offer a BI dashboard that highlights any stock price anomalies.
    a. Set up Kinesis Firehose to stream data to Redshift
    b. Create a BI dashboard using QS that has Redshift as its data source

135. Users of Redshift cluster include analysts running complex, long running queries as well as automated systems running short, transactional read only queries. During peak usage times, your transactional queries are timing out while the analysts are running their jobs. The simplest solution:
    a. Use Short Query Acceleration (SQA)
136. A large news website needs to produce personalized recommendations for articles to its readers by training a ML model on a daily basis using historical click data. The influx of this data is daily constant, except during major elections when tragedy to the site spikes considerably. The most cost effective and reliable solution would be:
    a. Publish click data into S3 using Kinesis Firehose and process the data nightly using Spark and MLLib on Spot instances in an EMR cluster.
    b. Publish the models results to DynamoDB for producing recommendations in real time
137. Create an EMR cluster that processes data in several MapReduce steps. You are working against the data in S3 using EMRFS but the network costs are extremely high as the processes write back temporary data to S# before reading it. You are tasked with optimizing the process and bringing the cost down:
    a. Add a preliminary step that uses S3DistCP command
    b. Copies data from S3 to HDFS and then ensure that data is processed locally by EMR cluster MapReduce job
    c. Upon completion, use S3DictCP to push back the final result data to S3
138. Loading several hundred GBs of data everyday from S3 into Redshift, which is stored in a single file. This data is prohibitively slow, the best approach for optimizing this loading:
    a. Split data into files between 1 to 125 MB (after compression) and specify GZIP from a single COPY command
139. Required to maintain a real time replica of Redshift data warehouse across multiple AZs. One approach is to spin up separate Redshift clusters in multiple AZs, using Kinesis to simultaneously write data into each cluster.
    a. Use Route53 to direct your analytics tools to the nearest cluster when querying your data
140. A financial services company wants to back up its encrypted data warehouse in Redshift daily to a different region. The simplest solution that preserves encryption in transit and at rest:
    a. Configure Redshift to automatically copy snapshots to another region
    b. Use KMS Customer Master Key in the destination region
141. Process data coming from IoT devices takes about 2 minutes per data point. Want to be able to scale in terms of the number of processes that will consume that data, based on the load receiving and no ordering constraints are required:
    a. Define an IoT results action to send data to SQS and consume the dta with EC2 instances in an Auto Scaling Group
142. Want to design an app that is able to sustain hundreds of TBs of data in a database that will get low latency on reads and won’t require you to manage scaling:
    a. Use DynamoDB

143. Company collects data from various sources into Kinesis and the stream is delivered using Kinesis Firehose into S3. Once in S3, your DS team uses Athena to query the most recent data and usage has shown that after a month, the team queries the data less and finally doesn't use it after 2 months. For regulatory reasons, it is required to keep all the data for 5 years and no one should be able to delete it. To do so:
    a. Implement a Glacier Vault Lock Policy
    b. Create a lifecycle rule to migrate all the data to S3 IA after 3 days and delete the data after 60 days
    c. Create a replicate of all the source data into Glacier from the first day
144. A system is created that recommends items similar to other items on an ecommerce website by training a recommender system using Mahout on an EMR cluster. A performant means to vend the resulting table of similar items for any given item to the website at high transaction rates:
    a. Publish the data into HBase
145. A real estate company wants to display interactive charts on their public facing website summarizing their prior months sales activity. Two solutions that would provide this capability in a scalable and inexpensive manner:
    a. Publish data in CSV format to Cloudfront via S3 and use Highcharts to visualize data on the web
    b. Publish data in CSV format to Cloudfront via S3 and use d3.js to visualize data on the web
146. Deploying your web servers (in a private subnet) from a Load Balancer (in a public subnet). For security reasons, the private subnet does not have access to the Internet and you want to ensure the web servers within this subnet have access to DynamoDB. To do so:
    a. Provision a VPC Endpoint Gateway
147. A company wishes to copy 500 GB of data from their Redshift cluster into a RDS postgreSQL database, in order to have both column and row based data stores available. The redshift cluster will continue to receive large amounts of new data everyday that must be kept in sync with the RDS database. The most efficient strategy:
    a. Copy the data using the dblink function into the PostgreSQL: tables
148. To monitor your clusters performances as a whole and as individual nodes: use Ganglia (EMR tool)

149. Three ways in which EMR integrates HBase with S3:
    a. Storage of HBase StoreFiles and metadata on S3
    b. Snapshots of HBase to S3
    c. HBase read replicas on S3
150. A hospital monitoring sensor data from heart monitors wishes to raise immediate alarms of an anomaly in any individual's heart rate. The architecture that meets these requirements in a scalable manner:
    a. Publish sensor data into a Kinesis Data Stream
    b. Create a Kinesis Analytics app using random_cut_forest to detect anomalies
    c. When an anomaly is detected, use Lambda to alarm SNS
151. Need to add more nodes to Redshift cluster and change the node types in the process. The process that allows you to do this while minimizing downtime for both reads and writes:
    a. Classic Resize with snapshot, restore, resize
152. An online retailer's website is a storefront for millions of products. You recently ran a big sale on one specific electronic and encountered a ProvisionedThroughput exception. Want to ensure you can survive an upcoming sales 3x as big:
    a. Enable DynamoDB DAX
153. Want to use Redshift Spectrum to analyze data in an S3 bucket that is in a different account than Redshift Spectrum. To authorize access between Spectrum and S3 across accounts:
    a. Add a policy to the S3 bucket allowing S3 GET and LIST operations for an IAM role for Spectrum on Redshift account
154. The security mechanisms that are supported by EMR:
    a. SSE-KMS, KMS Encryption, LUKs encryption
155. An app hosted on AWS needs to process game results immediately in real time and later perform analytics on the same game results in the order they came at the end of business hours:
    a. Use Kinesis Data Streams
156. Company has data from a variety of sources, including Microsoft Excel spreadsheets stored in S3, log data stored in S3 data lake and structured data stored in Redshift. The simplest solution for providing interactive dashboards that span data:
    a. Use QuickSight directly on top of Excel, S3, Redshift data
157. You are tasked with using Hive on Elastic MapReduce to analyze data that is currently stored in a large relational database. One approach to meet this requirement is:
    a. Use Apache Sqoop on EMR cluster to copy the data into HDFS
158. An app processes sensor data in real time by publishing it to Kinesis Data Streams, which sends data to a Lambda that processes it and feeds it to DynamoDB. During peak usage, it has been observed that some data is lost. You determined that you have sufficient capacity allocated for Kinesis shards and DynamoDB reads / writes. Two possible solutions:
    a. Increase Lambas timeout va;ue
    b. Process data in smaller batches to avoid hitting Lambdas timeout

159. As part of an effort to limit cost and maintain under control the size of your DynamoDB table, the AWS AM wants to ensure that old data is deleted in DynamoDB after 1 month. To do so with as little maintenance as possible and without impacting the current read and write operations:
    a. Enable DynamoDB TTL and add a TTL column
160. A MapReduce job on an EMR cluster needs to process data that is currently stored in very large, compressed files in HDFS (limits the cluster's ability to distribute processing). Two solutions that help MapReduce job operate most efficiently:
    a. Uncompress the data and split it into 128 MB chunks
    b. Convert the file in AVRO format
161. Three ways in which EMR integrates Pig with S3:
    a. Directly writing HCatalog tables in S3
    b. Submitting work from EMR console using Pig scripts stored in S3
    c. Loading custom JAR files from S3 with the REGISTER command
162. A bank is regularly uploading 100 MB files to S3 and analyzed by Athena. Some of the recent uploads have been corrupted and made a critical data job fail. Want a stronger guarantee that uploads are done successfully and that files have the same content on premise as on S3 with minimal cost:
    a. Use the S3 ETag and compare to the local MD5 hash
163. An S3 bucket can be read by the entire organization. For security reasons, I want the data to be encrypted and want to define a strategy in which users can only read the data which they are allowed to decrypt, which may be a different partial set of objects within the bucket for each user. To achieve that:
    a. Use SSE-KMS to encrypt the files
164. Dealing with PII datasets and wanting to leverage Kinesis Data Streams for your pub sub solution. Regulators imposed the constraint that the data must be encrypted end to end using an internal key management system:
    a. Implement a custom encryption code in the Kinesis Producer Library (KPL)
165. An organization has a large body of web server logs stored on S3 that it wants to analyze using Athena. Most queries are operational in nature and are limited to a single day's logs. To prepare data for minimal costs and performant queries:
    a. Convert the data into Parquet format that is compressed with Snappy
    b. Store in a directory structure of year=XXXX/month=XX/day=XX in S3
166. Data lake in S3 for user transaction data is stored in CSV format. To save costs, data older than 1 year is archived to Glacier. You received a request from your legal team to retrieve any records for a specific user ID going back 7 years and they need it today. The most cost effective way to fulfill this request:
    a. Use Glacier Select to run a query against the Glacier for this user ID and merge it with the same query for S3 Select
167. As an e-commerce retailer, you want to onboard clickstream data into Kinesis from your web servers JAva apps. You want to ensure that a retry mechanism is in place, as well as good batching capability and asynchronous mode. You want to collect server logs with the same constraints. It is recommended to:
    a. Use the Kinesis Producer Library to send the clickstream and the Kinesis Agent To collect Server Logs
168. As part of your app development, you want users to have Row Level Security. The app will be deployed on web servers and the users of the app should be able to see their amazon.com accounts. For the database and security:
    a. Enable Web Identity federation
    b. Use DynamoDB to reference ${www.amazon.com:user_id} in attached IAM policy
169. Have an ETL process that collects data from different sources and third party providers, would like to ensure that data is loaded into Redshift once all the parts from all the providers related to one specific job have been gathered. This process can happen over the course of one hour to one day. The least costly way of doing so:
    a. Create a Lambda that responds to S3 Upload Events and checks if all the parts are there before uploading to Redshift
170. A financial services company has a large, secure data lake stored in S3. They want to ana;yze this data using a variety of tools, including Apache Hive, Athena, Hive, Quicksight. To connect their DA tools to minimize cost and development work:
    a. Run a Glue Crawler on the data lake to popular a Glue Data Catalog
    b. Share the Glue Data Catalog as metadata repository between Athena, Redshift, Hive, QuickSight
171. New data arrives in S3 on an irregu;ar schedule to import into an ElasticSearch c;uster as it is received. The raw data in S3 requires custom parsing before it is loaded into Elasticsearch. To minimize ongoing maintenance / maximize scalability:
    a. Use Lambda to respond to event triggers from S3
    b. Stream data from S3 into ElasticSearch as it is received
172. Working for a data warehouse company that uses Redshift c;suter. For security reasons, it is required that VPC Flow Logs be analyzed by Athena to monitor all copy and and unload traffic of the c;suter that moves in and out of the VPC:
    a. Use Enhanced VPC Routing to force Redshift to use the VPC for all copy and unload commands (helps ensure that al traffic appears on the VPC Flow Logs)