AWS Data Analytics & Serverless Q&A

1. Difference between AWS Regions, Availability Zones, and Edge Locations

AWS Regions are geographical areas that contain multiple isolated data centers. Availability Zones (AZs) are individual data centers within a region, designed for fault isolation. Edge Locations are global endpoints that serve cached content to users with low latency.

Importance: Crucial for high availability and low-latency data analytics applications.

2. AWS CLI Command to List All AWS Regions

Command:

aws ec2 describe-regions --all-regions --query "Regions[*].RegionName" --output table

This returns all AWS regions in a table format.

3. Create IAM User with Least Privilege Access to S3

IAM Policy:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": ["s3:ListBucket", "s3:GetObject", "s3:PutObject"],
      "Resource": [
        "arn:aws:s3:::your-bucket-name",
        "arn:aws:s3:::your-bucket-name/*"
      ]
    }
  ]
}
```

4. Compare Amazon S3 Storage Classes

Standard: Active data

Intelligent-Tiering: Automatic tiering

Glacier: Archival storage

Use Standard for active datasets, Intelligent-Tiering for unknown access patterns, and Glacier for long-term storage.

5. Create S3 Bucket, Upload Dataset & Enable Versioning

Use AWS CLI to create a bucket, enable versioning, and upload files. Then list versions using:

aws s3api list-object-versions --bucket my-data-bucket --prefix sample.csv

6. Lifecycle Policy to Move to Glacier after 30 Days and Delete after 90

JSON Policy:

```
{
  "Rules": [
    {
      "ID": "MoveToGlacierAndExpire",
      "Status": "Enabled",
      "Transitions": [{"Days": 30, "StorageClass": "GLACIER"}],
      "Expiration": {"Days": 90}
    }
  ]
}
```

7. Compare RDS, DynamoDB, and Redshift

RDS: For structured relational data

DynamoDB: NoSQL, real-time lookups

Redshift: Analytical processing over large datasets

Use each depending on the stage and structure of your data.

AWS Data Analytics & Serverless Q&A

8. Lambda Function to Add Records on S3 Upload

Python Lambda reads S3 trigger and writes to DynamoDB. It logs file name and status into the table FileLogTable.

9. What is Serverless Computing (AWS Lambda)?

Lambda lets you run code without managing servers.

Pros: No infrastructure, auto-scaling

Cons: Timeout limits, cold starts

Use in data pipelines for ETL and event-driven tasks.

10. Lambda Function Triggered by S3 Upload Logging to CloudWatch

Python Lambda logs file name, size, and timestamp from S3 event. Output is logged in CloudWatch logs.

11. Glue Job to Convert CSV to Parquet

Use Glue to crawl S3, create a Data Catalog table, and run a PySpark job to write data in Parquet format to another S3 location.

12. Kinesis Services Comparison

Kinesis Data Streams: Real-time ingestion

Kinesis Firehose: Delivery to S3/Redshift

Kinesis Data Analytics: SQL on streaming data

Each used for real-time processing with different destinations.

13. Columnar Storage in Redshift

Stores data by columns rather than rows. Benefits: Faster analytical queries, reduced I/O, better compression.

14. Load CSV into Redshift using COPY

COPY sales_data FROM 's3://bucket/file.csv' IAM_ROLE 'arn:aws:role' FORMAT AS CSV IGNOREHEADER 1;

AWS Data Analytics & Serverless Q&A

15. Glue Catalog in Athena & Schema-on-Read

Glue Data Catalog holds metadata. Athena uses schema-on-read to query raw data without transforming it upfront.

16. Athena Table from S3 using Glue Catalog

Use a Glue Crawler to create a table, then run SQL in Athena to analyze the data using standard SQL syntax.

17. Amazon QuickSight for BI

Serverless BI with SPICE (in-memory engine) and embedded dashboards. Great for scalable, interactive analytics.

18. Connect QuickSight to Athena or Redshift

Build visuals with calculated fields (e.g., TotalRevenue = price * quantity) and filters (e.g., product = 'Widget A').

19. CloudWatch vs CloudTrail

CloudWatch: Logs and metrics for monitoring.

CloudTrail: Tracks user actions for auditing.

Both help in observability of analytics pipelines.

20. End-to-End AWS Data Analytics Pipeline

S3 -> Lambda -> Glue -> Athena -> QuickSight

Each stage serves ingestion, transformation, querying, and visualization using fully managed, serverless AWS services.