Advanced Amazon S3 and Athena

Uday Manchanda

September 7, 2021

1 S3

1.1 S3 MFA Delete

- Enable versioning on bucket
- Need MFA to delete an object version or suspend versioning
- Can be enabled only via CLI

1.2 S3 Default Encryption

- One way to "force encryption" is to use a bucket policy and refuse any API call to PUT an S3 object w/out encryption headers
- Another way is via "default encryption" option
- Can be enabled in the console

1.3 Access Logs

- May need to log all access to s3 buckets (athena)
- Could send all logs to a specific logging bucket
- Do not send your logging bucket to be the monitored bucket

1.4 Replication

- CRR = Cross region replication
- SRR = Same region replication
- Bucket in region A, asynchronously replicate in region B
- Buckets can be in different accounts

- Use cases: compliance, lower latency, replication across accounts
- After activating, only new objects are replicated
- No chaining of replication

1.5 Pre Signed URLs

- Can generate pre-signed URLs using SDK or CLI
- valid for default of 3600 seconds, can change
- EX: allow only logged-in users to download a premium video on S3, allow temporarily a user to upload a file to a precise location in our bucket

1.6 S3 Storage Classes

- Amazon S3 standard
  - General purpose
  - High durability
  - Use cases: big data analytics, mobile/gaming apps, content distribution
- Amazon S3 Standard-Infrequent Access (IA)
  - Suitable for data less frequently accessed, but requires rapid access when needed
  - Data store for DR
- Amazon S3 One Zone-Infrequent Access
  - Same as previous but data is stored in a single AZ
- Amazon S3 Intelligent Tiering
  - Automatically moves objects between 2 access tiers based on changing access patterns
- Amazon Glacier
  - Low cost object storage meant for archiving/backup
  - Data is retained for the long term
  - Alternative to on premise magnetic tapes
  - each item in glacier is called an archive
  - archives are stored in vaults
  - min storage duration is 90 days
- Amazon Glacier Deep Archive
  - Three retrieval options for glacier: expedited, standard, bulk
  - Super long term storage (min 180 days)

1.7 S3 Lifecycle Rules

- Can transition objects between storage classes
- Moving objects can be automated using a lifecycle configuration
- Transition actions, expiration actions
- Can use analytics to help determine when to transition objects

1.8 S3 Performance

- automatically scale to high request rates, latency
- how to optimize performance
  - multi part upload - parallelize uploads for large files
  - s3 transfer acceleration - increase transfer speed by transferring files to an AWS edge location which will forward the data to the s3 bucket in the target region

1.9 S3+Glacier Select

- retrieve less data using SQL, can filter by rows and columns
- less network transfer

1.10 S3 Requester Pays

- In general bucket owners pay for all s3 storage and data transfer costs
- with requester pays buckets, the requester pays for these costs instead

2 Athena

- Serverless service to perform analytics against S3 files
- Uses SQL to query the files
- Charged per query and amount of data scanned
- Analyze data DIRECTLY in S3

3 Quiz

- Question
  - you are looking to build an index of your files in s3 using RDS postgres
  - to build this index it is necessary to read the first 250 bytes of each object in s3
  - which contains some metadata about the content of the file itself
  - there are over 100,000 files in your s3 bucket
  - how can you build this index efficiently
- Answer
  - Create application that will traverse s3 bucket
  - issue byte range fetch for the first 250 bytes
  - store that info in RDS