AWS ML - Data Engineering

1. Amazon S3

Amazon S3 Storage

- Buckets must have a globally unique name
- Objects (files) have a Key. **The key is the FULLpath**
- Max object size is **5TB**
- supports any file format : CSV, JSON, Parquet, ORC, Avro, Protobuf

- **Object Tags** (key / value pair –up to 10) –useful for security / lifecycle
- Data Partitioning : Pattern for speeding up range queries
- AZ : Available zones

S3 Storage Classes

- Amazon S3 **Standard** -General Purpose :
  - Used for frequently accessed data
- Amazon S3 **Standard IA** -Infrequent Access
  - Use cases: Disaster Recovery, backups
- Amazon S3 **Intelligent Tiering :**
  - Moves objects automatically between Access Tiers based on usage
  - There are no retrieval charges
- Amazon S3 **One Zone IA** -Infrequent Access
  - Use Cases: Storing secondary backup copies of on-premise data, or data you can recreate
- Amazon S3 Glacier **Instant Retrieval :** great for data accessed once a quarter
- Amazon S3 Glacier **Flexible Retrieval**
- Amazon S3 Glacier **Deep Archive :** Minimum storage duration of 180 days
- Glacier :
  - Low-cost object storage meant for archiving / backup
  - Pricing: price for storage + object retrieval cost
- You can move between classes **manually** or using **S3 Lifecycle configurations.**
- 99.999999999% (11 9's) **durability** for **all storage classes**

Amazon S3 – Lifecycle Rules :

- can be created for for a certain prefix/certain objects Tags.
- **Transition Actions :** configure objects to transition to another storage class.
- **Expiration Actions :**
  - Can be used to delete old versions of files (if versioning is enabled)
  - Can be used to delete incomplete Multi-Part uploads.

Amazon S3 Analytics - Storage Class Analysis:

- Help you decide when to transition objects to the right storage class
- Does NOT work for One-Zone IA or Glacier

Amazon S3 – Security

- **User-Based**
  - **IAM Policies** – which API calls should be allowed for a specific user from IAM
- **Resource-Based**
  - **Bucket Policies** – bucket wide rules from the S3 console - allows cross account
  - **Object Access Control List** (ACL) – finer grain (can be disabled)
  - **Bucket Access Control List** (ACL) – less common (can be disabled)
- Note: an IAM principal can access an S3 object if • The user IAM permissions ALLOW it OR the resource policy ALLOWS it • AND there's no explicit DENY

Amazon S3 – Object Encryption 2 categories, 4 methods

- **Server-Side Encryption (SSE)**
  - **Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3)** : Enabled by Default : Encrypts S3 objects using keys handled, managed, and owned by AWS. Encryption type is AES-256
  - **Server-Side Encryption with KMS Keys stored in AWS KMS (SSE-KMS) :** Leverage AWS Key Management Service to manage encryption keys. KMS advantages: user control + audit key usage using CloudTrail

```
When you upload, it calls the **GenerateDataKey KMS API**
 When you download, it calls **the Decrypt KMS API**

```

  - **Server-Side Encryption with Customer-Provided Keys (SSE-C) :** When you want to manage your own encryption keys.
    - Amazon S3 does **NOT store** the encryption key you provide
    - **HTTPS must be used**
- **Client-Side Encryption:**
  - Use client libraries such as **Amazon S3 Client-Side Encryption Library**
  - Customer fully manages the keys and encryption cycle (encryption and decryption)

Amazon S3 – Encryption in transit (SSL/TLS) : Encryption in flight is also called SSL/TLS

Amazon S3 exposes two endpoints:

- HTTP Endpoint – non encrypted
- HTTPS Endpoint – encryption in flight

Note: **Bucket Policies are evaluated before "Default Encryption"**

VPC Endpoint gateway for S3

- **Privately** access your S3 bucket without going through the public internet
- How does Amazon VPC route traffic between subnets by default? **Using Route Tables**
- **Internet Gateway** takes care of communication between instances in Amazon VPC & Internet.
- **Network Access Control Lists (NACLs)** takes care of security.
- **Security Groups** take care of Incoming traffic.

2 AWS Kinesis (Real Time)

- is a real-time data streaming platform.
- It enables the collection, processing, and analysis of large streams of data in real-time
- 4 services:

Amazon Kinesis Data Streams (KDS)

- Allows you to collect and process large streams of data records in real-time.
- "To stream data" se traduit par la notion de transférer des flux de données en continu entre un producteur et un consommateur en temps réel.
- **Use Cases**: Website activity tracking, log collection, real-time event processing
- Once data is inserted in Kinesis, it can't be deleted (immutability)
- need capacity planning,
- Data Storage for 1 to 365 days, **replay capability**, **multi consumers**
- Automatic scaling with On-demand Mode
- **Features:**

- **Partitioning**: Data is partitioned into **shards** (fragments), allowing for parallel processing.
- **Retention Period**: By default, data is retained for 24 hours, extendable up to 365 days
- **Capacity Modes:**
  - **Provisioned mode:** You choose the number of shards provisioned, You pay per shard provisioned per hour
  - **On-demand mode:** Scales automatically based on observed throughput peak during the last 30 days. You pay per stream per hour & data in/out per GB
- **Limits:**
  - **Producer** : **1MB/s at write per shard or 1000 messages/s (maximum throughput of a single shard)**
  - **Consumer Classic :**
    - 2 MB/s at read per shard
    - 5 API calls per second PER SHARD across all consumers
- The more capacity you need in your streams, you need to add shards.

Amazon Kinesis Data Analytics

- **SQL transformations on streaming data**
- Enables real-time data stream analysis **using SQL or Flink** to gain real-time insights.
- **Use Cases**: streaming ETL, Anomaly detection, real-time dashboards, data quality monitoring, etc
- Pay only for resources consumed.
- Serverless; scales automatically
- Use IAM permissions to access streaming source and destination(s)
- Lambda can be used for pre-processing
- Instead of using SQL, you can develop your own Flink application from scratch and load it into MSAF via S3
- Destinations : S3 and Redshift

Amazon Kinesis Data Firehose:

- **Delivery/ Ingestion service**
- Fully Managed Service, no administration
- **Near Real Time data Ingestion :** load near real-time data streams into destinations such as:
  - **Amazon S3**
  - **Amazon Redshift**
  - **Amazon Elasticsearch Service**
  - **Splunk**
  - **Amazon OpenSearch**
  - **Snowflake**
  - **Custom HTTP endpoint**
- **Use Cases**: Log collection, real-time analytics, real-time data storage
- Can compress records when target is Amazon S3 (GZIP, ZIP, and SNAPPY)
- Pay for the amount of data going through Firehose
- Serverless data transformations with Lambda (small compute service that allows you to run functions in the cloud without provisioning servers)
- **Features :**
  - **Data Transformation**: Data can be transformed using AWS Lambda before being loaded into the destination.
  - convert record format to :
    - Apache Parquet

        - Apache ORC
    - **Automatic Scaling**: Automatically adjusts capacity to match the data throughput.

Amazon Kinesis Video Streams

- Enables you to stream, store, and analyze video streams.
- **Use Cases**: Video surveillance, video analytics, live streaming, etc.
- connecting video data from cameras to the AWS cloud for further processing with services such as Rekognition, SageMaker, or EC2.
- **Features**:
    - **Video Ingestion**: Continuous ingestion of video from cameras, mobile devices, drones, etc.
    - **Storage and Playback**: Store video for later playback and real-time analysis.
- Keep data for 1 hour to 10 years
- Video playback capability
- One stream per device

Machine learning algorithms on Kinesis:

1. RANDOM_CUT_FOREST: ever changing model
    - SQL function used for anomaly detection on numeric columns in a stream
    - Uses **recent** history to compute model
2. Hotspots
    - locate and return information about relatively dense regions in your data

3. Glue

AWS Glue is a service that helps you move and transform data from various sources to a single place for analysis.

Glue Data Catalog & Crawlers: Metadata repositories for schemas and datasets in your account

Glue Data Catalog :

a unified metadata repository for all your data assets, regardless of where they are stored.

Glue Crawlers:

- Automatically scans your data sources to update the Data Catalog with the latest data structure.
- Crawlers work for: S3, Amazon Redshift, Amazon RDS
- Need an IAM role / credentials to access the data stores
- will extract partitions based on how your S3 data is organized

Glue ETL:

- Transform data, Clean Data, Enrich Data (before doing analysis)
- Underlying platform for Glue ETL is : a **serverless Apache Spark Cluster**
- ETL Jobs as Spark programs, run on a **serverless Spark Cluster**
- **Bundled Transformations:**
    - **DropFields, DropNullFields** – remove (null) fields
    - **Filter** – specify a function to filter records
    - **Join** – to enrich data

    - **Map** - add fields, delete fields, perform external lookups
- **Machine Learning Transformations:**
    - **FindMatches ML:** identify duplicate or matching records in your dataset, even when the records do not have a common unique identifier and no fields match exactly.

Glue Scheduler

to schedule the jobs

Glue Triggers

to automate job runs based on “events”

AWS Glue DataBrew

- Allows you to clean and normalize data **without writing any code**
- Reduces ML and analytics data preparation time by up to 80%
- Data sources include S3, Redshift, Aurora, Glue Data Catalog…
- +250 ready-made transformations to automate tasks : Filtering anomalies, data conversion, correct invalid values…

4. AMAZON Athena

is an interactive **query** service that makes easy to analyse **instantly** data using **SQL**

- Supports Multiple Data Formats
- **Serverless**: Athena is a fully managed service,

Serverless interactive queries of S3 data (SQL)

no need to load data, it stays in S3

supports many data formats, structured and unstructured data

Pay as you go

Applications :

- Analyse logs from CloudTrail/ CloudFront/ VPC
- Run queries from Jupyter, Zeppelin, RStudio notebooks

AWS Glue can transform data or convert it into column formats to optimize the cost of your Athena queries.

Converting data to columnar formats (ORC, Parquet) save a lot of money and get better performance

AWS Data Stores

Redshift:

- Data Warehousing, SQL **analytics** (**OLAP** - Online analytical processing)
- **Columnar-based**
- you can run massively parallel SQL queries to perform some analytics
- Load data from S3 to Redshift or Use **Redshift Spectrum** to query data directly in S3 (no loading)
- **Redshift Spectrum:** Redshift on data in S3 (without the need to load it first in Redshift)
- you need to provision in advance, it’s like an entire database and then you run your SQL analytics on it

- The primary purpose of the leader node in Amazon Redshift is to distribute queries and manage query optimization.

RDS, Aurora

- Relational Store, SQL (**OLTP** - Online Transaction Processing)
- **row-based**
- **Must provision servers** in advance

DynamoDB

- **NoSQL** (not only SQL) data store
- **serverless**
- provision read/write capacity
- Useful to store a machine learning model served by your application (your model output maybe stored into dynamodb)

OpenSearch (previously ElasticSearch):

- Indexing of data
- Search amongst data points
- Clickstream Analytics

ElastiCache:

- Caching mechanism : data can easily and fastly accessed
- Not really used for Machine Learning

AWS Data Pipeline Features

- a **Orchestration** service to move data from one plcae to another →Orchestration of ETL jobs between RDS, DynamoDB, S3. Runs on EC2 instances
- manages task dependencies
- For example, if you need to move data from RDS to S3, AWS Data Pipeline will create a EC instance or many and these EC2 instances would be tasked for that

AWS Data Pipeline vs Glue : both are ETL services

- Glue:
  - Glue ETL - you run **Apache Spark** code, Scala or Python based, focus on the ETL
  - Glue ETL - you do not worry about configuring or managing the resources (**all ressources belong to AWS)**
  - Data Catalog helps you to make the data available to Athena or Redshift Spectrum
- Data Pipeline:
  - Orchestration service (it does not run the stuff for you)
  - You have more control over the environment, the compute resources that run code, & the code
  - Allows you to get access to EC2 or EMR instances (**creates resources in your own account**)
- **Which is a fully managed ETL service for categorizing, cleaning, enriching, and moving your data?**

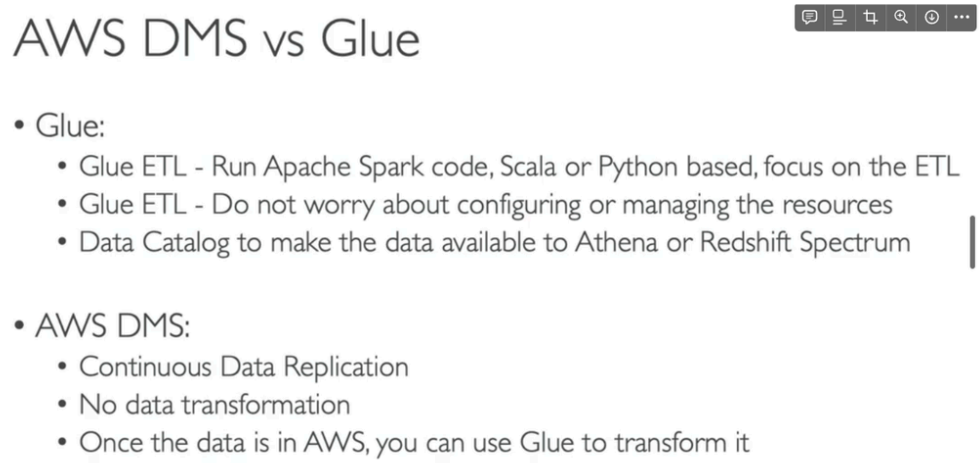
  AWS Glue

AWS Batch

- A service that allows you to run batch jobs as Docker images - not just for data, manages EC2 instances for you
- for any computing job **(not only ETL like Glue)**
- Ressources are created in your account (not like Glue)
- You don’t need to provision the instances → Dynamic provisionning
- Optimal quantity and type based on volume and requirements
- No need to manage clusters, **fully serverless (like Glue)**
- You just pay for the underlying EC2 instances
- Schedule Batch Jobs using CloudWatch Events
- Orchestrate Batch Jobs using AWS Step Functions

DMS – Database Migration Service

- Quickly and securely **migrate databases to AWS**, resilient, self healing
- Supports: • Homogeneous migrations: ex Oracle to Oracle • Heterogeneous migrations: ex Microsoft SQL Server to Aurora
- No data transformation : Once the data is in AWS, you can use Glue to transform it
-



AWS Step Functions

- Orchestrate and design **workflows.**
- Provides advanced Error Handling and Retry mechanism outside the code.
- You can know how every service reacted

Course: AWS Certified Machine Learning Specialty 2024 - Hands On! | Udemy

Other random notions

AWS DataSync

- getting data from inside company to AWS and probably to S3
- A DataSync Agent is deployed as a VM and connects to your internal storage • NFS, SMB, HDFS

MQTT

- IOT : Think of it as how lots of sensor data might get transferred to your machine learning model
- The AWS IoT Device SDK can connect via MQTT

Other Notions

- EMR: Managed Hadoop Clusters : Analyzing and processing big data
  - The --output argument is used to define the location to store output data in Amazon S3 or HDFS for an EMR step
- Quicksight: Visualization Tool
- Rekognition: ML Service
- SageMaker: ML Service
- DeepLens: camera by Amazon
- Athena: Serverless Query of your data

Quiz

What is the primary purpose of CloudWatch Log Groups? Grouping log streams for easy management

Cluster is a group of EC2 instances