Data and Analytics

Athena

Query service that let you analyze data stored in S3. It uses SQL language to do the query.

You will be putting data into S3 then you will query it with Athena.

You pay per TB of data scanned. No need to pay for any server since it is serverless.

It is commonly used with Quicksights for reporting/dashboards.

Use case: Business intelligence / analytics / reporting and analyze logs

Serverless SQL to analyze S3 use Athena!

Improving performance

To improve Athena for performance you want to scan less data. Apache Parquet or ORC is recommended, you want to use Glue to convert your data. So that you can get column data and not all the data.

Do some compression for smaller retrievals/

Partition your S3 so that it is easier for querying virtual columns.

You only want to get subset of the column so you will be paying less.

Use larger files since it is easier to scan > lots of smaller files

Federated query

Allows you to run SQl queries across data stored in relational, non-relational, object, and custom data source. **So basically let you run Athena on any type of databases even on-premise database.**

You need a Data Source Connectors that runs on AWS Lambda to run Federated Queries.

Redshift

It is a database but also do analytics. The database is based on PostgresSQL. For data warehouse and analytics.

Load your data into Redshift and then do analytics. Column of storage data and then do parallel query engine.

SQL interface for performing queries.

Comparing with Athena, you have to load data into Redshift first, then it will have much faster query and join compared to Athena. It has indexes to have high performance. If many query, joints needed Redshift is better than Athena.

Redshift Cluster

Provision the node size in advance, can use reserved instances for cost saving

Has leader node for query planning and result aggregation.

Compute node: Actually carry out the queries

Snapshots and disaster recovery

Redshift can have multi-AZ for some clusters.

Snapshots are point in time backups of a cluster stored in S3, you can restore the snapshot to a new cluster. Snapshots can be taken automatically or you can do manual snapshot retained until you delete it.

You can also automatically copy snapshot of a cluster to another AWS region to do disaster recovery.

Inserting data into Redshift

You can insert data into Redshift via Kinesis data firehose (which will deposit it into S3 then it will issue a copy from S3 to redshift)

You can also copy the data from S3 from Redshift by issuing a copy command with the correct IAM role. Do it with Enhanced VPC Routing this so that the traffic is moving via VPCs and not via public internet because S3 is publicly accessible.

Or you can write from you EC2 instances into Redshift, but do it in large batch much more efficient.

Redshift spectrum

This is how you can query data that is already in S3 without loading it into Redshift. How do we do it? We have a redshift cluster available that's a must, then you will submit the query from your cluster, that query will reach the Redshift spectrum nodes (they sit in front of your S3 and then query the result, there are thousands of them so it is very efficient), then it will return you the data to the one that asked for the query.

You can leverage the nodes in Redshift spectrum rather than your own cluster to perform much more efficient query.

OpenSearch

Successor to ElasticSearch

DynamoDB you can only query by primary key or indexes, but with OpenSearch you can search in any field. You would use OpenSearch as a complement to another database. And you can also do analytics.

You will need to back it up with a cluster since it is not serverless.

The query language is not SQL it has it's own language.

Common pattern with OpenSearch

You have DynamoDB containing you data, you sent changes to DynamoDB stream with Lambda function reacting to it, and lambda insert it into OpenSearch. Finally you can then search using OpenSearch and retrieve the item from DynamoDB.

You can also sent CloudWatch log to lambda then sent it to OpenSearch to do some searching.

Kinesis Data Stream sent data to Kinesis Data firehose and optionally transform the data, then near real time sent data to OpenSearch.

Or you can sent Kinesis Data stream sent it to Lambda function to read it and write it into OpenSearch.

EMR

Elastic MapReduce. Create a Hadoop clusters for doing big data analyze.

The hadoop cluster will be hundreds of EC2 instances. EMR comes with lots of tools for big data scientist.

Apache Spark, HBase, presto, Flink. EMR helps you setting up with those tools for you.

Use case: data processing, machine learning, web indexing, big data

Master node: Manages the cluster, coordinate, and manage health, must be long running

Core node: Runs tasks and store data

Task node: Just to run tasks, usually spot instances

Purchasing options

On-demand: reliable, predictable, won't be terminated

Reserved: cost saving, used for master node and core node

Spot instances: Used for task nodes

QuickSight

Serverless machine learning powered business intelligence service to create interactive dashboards.

Fast, automatically salable, embedded, with per-session pricing.

Uses in-memory computation using SPICE engine, only if the data is imported into QuickSight. Doesn't work if you don't import the data.

Column-level security to prevent others seeing some columns.

You can use QuickSight with RDS, Aurora, RedShift, Athena, S3, OpenSearch, TimeSearch.

Dashboard and Analysis

Users and Groups only exist in the QuickSight they are not IAM.

Dashboard is read-only snapshot of an analysis that you can share.

Dashboard is then shared with users or groups. Users can also see the underlying data.

Glue

Managed extract, transform, and load ETL service. Fully managed

used to prepare and transform data for analytics.

Help convert data to parquet format

Parquet is column data, much better for filtering with Athena.

Glue can be used to transform data to parquet format.

Glue data catalog

Data crawler connect to databases. Write metadata of the columns to Data Catalog, then it is used to perform ETL.

Other things

Glue job bookmarks: Prevent re-processing old data
Glue elastic views: Combine and replicate data across multiple data stores using SQL.
Glue databrew: Clean and normalize data
Glue studio: GUI to create, run and monitor ETL jobs
Glue streaming ETL: Process streaming data as well

Lake formation

Data lake is a central place to store your data so you can do analytics on it.

Lake formation is managed service that make it easy to set up data lake in days. It is actually backed by a S3 underneath.

Automate collecting, cleansing, moving, cataloging data.

You can combine structured and unstructured data in the data lake. You can also migrate storage S3, RDS, NoSQL in AWS all to Lake formation.

You can also have row and column level control on Lake formation, finer grain of security.

Sources

Data source can be from S3, RDS, Aurora, sent these data into Lake formation.

Consumer

Athena, Redshift, EMR can be used to perform analytics.

Centralized permissions

There are multiple places to manage security so it is a mess. Lake formation solves this, you have a one place to manage your security on row and column level because all data are sent to Lake formation. So you don't have to manage the security everywhere.

Amazon MSK (Managed streaming for Apache Kafka)

Kafka is an alternative to Kinesis, both allow you to stream data.

There are broker. Which is like shards in Data Stream

Producer will produce data and they will sent the data to the broker which then is replicated across other brokers.

Consumer will consume the data, process it and sent it to other places. Can be Kinesis Data Analytics for Apache Flink, Glue, Lambda, or you custom application running on ECS, EKS, or EC2.

MSK manages Apache Kafka on AWS for you.

Data is stored on EBS volumes for as long as you want.

You also have the option to run Apache Kafka serverless.

Data stream vs MSK

1 MB message size limit per shard in Data Stream. But MSK can have higher input limits.

Kinesis Data Stream you can remove shards, but MSK you cannot only add.

Big data ingestion pipeline

IoT devices produces lots of data. The data are sent in real-time to Kinesis Data Streams.

Then you can forward data to a firehose then sent the data to S3 bucket.

You can then sent the deposit event to SQS queue, Lambda receives the event of the data being deposited into the S3, then will run Athena query and sent the query to S3 again.

Then you can use QuickSight to make dashboard on the data, or Redshift for actual analytics but Redshift isn't serverless!

Revision #4
Created 2023-02-24 20:08:04 UTC by Tamarine
Updated 2023-02-25 00:06:10 UTC by Tamarine